UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Miro Rogina

ZAGOTAVLJANJA PRIHODKA V ELEKTRODISTRIBUCIJI Z UPORABO PODATKOV PAMETNIH ŠTEVCEV

MAGISTRSKO DELO

Mentor: prof. dr. Marko Bajec

Ljubljana, 2016

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

Večna pot 113
1000 Ljubljana, Slovenija
telefon: 01 47 98 100
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Številka: 151-MAG-ISO/2016
Datum: 29. 02. 2016

Miro ROGINA, univ. dipl. inž. el.

L j u b l j a n a

Fakulteta za računalništvo in informatiko Univerze v Ljubljani izdaja naslednjo magistrsko nalogo

Naslov naloge: **Zagotavljanja prihodka v elektrodistribuciji z uporabo podatkov pametnih števcev**

Revenue assurance in electro distribution using smart meters data

Tematika naloge:

Identificirajte področja odtekanja prihodka v elektrodistribucijskem podjetju in predlagajte aktivnosti v procesu zagotavljanja prihodka (Revenue Assurance). Pri tem raziščite in predlagajte možnosti uporabe podatkov, ki jih elektrodistribucijsko podjetje zbere s pametnimi števci, z odbiranjem 15-minutnih porab energije pri odjemalcih. Z odkrivanjem znanja iz teh podatkov ter profiliranjem odjemalcev zastavite metode za odkrivanje kraje električne energije.

M e n t o r :

prof. dr. Marko Bajec

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

D e k a n :

prof. dr. Nikolaj Zimic

IZJAVA O AVTORSTVU ZAKLJUČNEGA DELA

Spodaj podpisani Miro Rogina, vpisna številka 63070484, avtor pisnega zaključnega dela študija z naslovom:

Zagotavljanja prihodka v elektrodistribuciji z uporabo podatkov pametnih števcev

IZJAVLJAM

1. da sem pisno zaključno delo študija izdelal samostojno pod mentorstvom
   prof. dr. Marka Bajca;
2. da je tiskana oblika pisnega zaključnega dela študija istovetna elektronski obliki pisnega zaključnega dela študija;
3. da sem pridobil/-a vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v pisnem zaključnem delu študija in jih v pisnem zaključnem delu študija jasno označil/-a;
4. da sem pri pripravi pisnega zaključnega dela študija ravnal/-a v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil/-a soglasje etične komisije;
5. soglašam, da se elektronska oblika pisnega zaključnega dela študija uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;
6. da na UL neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve avtorskega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja pisnega zaključnega dela študija na voljo javnosti na svetovnem spletu preko Repozitorija UL;
7. dovoljujem objavo svojih osebnih podatkov, ki so navedeni v pisnem zaključnem delu študija in tej izjavi, skupaj z objavo pisnega zaključnega dela študija.

V Ljubljani

Datum: 28.8.2016

Podpis študenta:

_____

Zahvala

Zahvaljujem se mentorju, prof. dr. Marku Bajcu, za strokovno usmerjanje in vsestransko pomoč pri pripravi magistrske naloge.

Zahvaljujem se podjetju Elektro Celje, d. d., ki mi je dovolilo uporabo podatkov in ponudilo infrastrukturo, na kateri sem te podatke za potrebe naloge obdeloval.

Posebna zahvala gre ženi Luciji, ki mi je z veliko mero potrpežljivosti, razumevanja in odrekanja vedno stala ob strani in me brezpogojno podpirala.

Hvala tudi sestri Anki za vso pomoč.

Vsebina

1 Uvod 3

1.1 Predstavitev problema 3

1.2 Namen in cilji naloge 4

2 Preprečevanje odtekanja prihodkov 5

2.1 Splošno o pristopu 'preprečevanje odtekanja prihodkov' 5

2.1.1 Stopnje zrelosti preprečevanja odtekanja prihodkov 6

2.1.2 Kakovost podatkov 6

2.1.3 Čiščenje podatkov 7

2.2 Preprečevanje odtekanja prihodkov v elektrodistribuciji 7

2.2.1 Organiziranost trga z električno energijo 7

2.2.2 Procesi, povezani z obračunavanjem omrežnine 9

2.2.3 Načini odtekanja prihodkov v elektrodistribuciji 11

2.2.4 Predlagan model za preprečevanje odtekanja prihodkov 13

2.3 Pametni števci 14

2.4 Dnevni obremenilni diagrami 16

3 Razvrščanje v skupine 19

3.1 Mere razdalj 19

3.2 Razvrščanje v skupine 20

3.3 Hierarhični postopki 20

3.4 Delitveni postopki 21

3.4.1 Metoda K-voditeljev (K-means) 22

3.5 Tehnike temelječe na modelih (verjetnostni pristop k razvrščanju) 23

3.5.1 Metoda maksimiranja pričakovanj (EM - Expectation-Maximization) 23

3.6 Klasifikacija - uvrščanje v predpisane skupine 26

3.6.1 Odločitvena drevesa 26

3.6.2 Povezovalna pravila 26

4 Zasnova platforme za podporo preprečevanju odtekanja prihodkov v elektrodistribucijskem podjetju 28

4.1 Opis infrastrukture 28

4.2 Priprava podatkovnega skladišča 30

4.2.1 Tabele dejstev 32
4.2.1.1 Dnevni diagrami odjemalcev 32
4.2.1.2 Dnevni diagrami pretokov v transformatorskih postajah 34
4.2.1.3 Obračunane količine energije 34
4.2.2 Dimenzija čas 36
4.2.3 Dimenzija merilno mesto 36
4.2.4 Dimenzija števec 37
4.2.5 Dimenzija pogodba o dostopu 37
4.2.6 Dimenzija omrežje 40
4.2.7 Dimenzija transformatorska postaja 42
4.2.8 Avtomatizacija procesa polnjenja skladišča – postopek ETL 43
4.2.9 Kakovost podatkov 46
4.2.9.1 Težava pri povezovanju števcev iz sistema AMI z eIS preko tovarniške številke 47
4.2.9.2 Težave z evidenco namestitev števcev v eIS 47
4.2.9.3 Težave z anomalijami pri pogodbi o dostopu 48
4.2.9.4 Pomanjkljivosti in napake pri opisovanju omrežja 48
4.2.9.5 Pomanjkljivosti in napake pri določanju mesta priključitve merilnega mesta na omrežje 49
4.3 Analitična struktura (OLAP) 49
4.4 Odkrivanje značilnih dnevnih diagramov s podatkovnim rudarjenjem 52
4.4.1 Izbira metode za razvrščanje daljinsko odbiranih merilnih mest v skupine tipične porabe 53
4.4.2 Primerjava tipičnih dnevnih diagramov porabe 59
4.4.3 Uvrščanje merilnih mest, ki nimajo podatkov o 15-minutnih porabah, v skupine 64
4.4.4 Avtomatizirano izvajanje podatkovnega rudarjenja in procesiranja OLAP kocke 69
4.5 Primeri poročil 71
5 Sklepne ugotovitve 77

Seznam uporabljenih kratic

kratica	Slovensko	Angleško
AMI	sistem naprednega merjenja	Advanced Metering Infrastructure
BI	poslovna inteligenca	Business Intelligence
BTP	baza tehničnih podatkov	
DMS	sistem za upravljanje distribucijskega sistema	Distribution Management System
DMX	razširitve za podatkovno rudarjenje	Data Mining Extensions
EDP	elektrodistribucijska podjetja v Sloveniji (Elektro Celje, Elektro Gorenjska, Elektro Ljubljana, Elektro Maribor, Elektro Primorska)	
eIS	obračunski sistem elektrodistribucij Slovenije	
EM	metoda za maksimiranje pričakovanj	Expectation-maximization
ET	enotna tarifa	
ETL	postopek zajemanja, preoblikovanja in polnjenja podatkov	Extract, Transform, Load
eTOM	standardiziran okvir poslovnih procesov v telekomunikacijah	Enhanced Telecom Operations Map
GIS	prostorski informacijski sistem	Geographic Information System
MDX	poizvedovalni jezik za večdimenzijske baze	Multidimensional Expressions
MT	manjša tarifa	
OLAP	sprotna analitična obdelava podatkov	Online Analytical Processing
RA	preprečevanje odtekanja prihodkov	Revenue Assurance
SQL	strukturirani povpraševalni jezik	Structured Query Language
SSAS	analitične storitve strežnika SQL	SQL Server Analysis Services
SSDT	podatkovna (razvojna) orodja strežnika SQL	SQL Server Data Tools
SSIS	integracijske storitve strežnika SQL	SQL Server Integration Services
SSMS	konzola za upravljanje strežnika SQL	SQL Server Management Studio
SSRS	poročevalske storitve strežnika SQL	SQL Server Reporting Services
VT	večja tarifa	
XMLA	razširljivi označevalni jezik za analize	Extensible Markup Language for Analysis

Povzetek

S povečevanjem kompleksnosti storitev naraščata tudi kompleksnost in heterogenost sistemov, ki jih ponudniki uporabljajo za merjenje in obračunavanje storitev. To povečuje možnosti za napake in posledično za izpad prihodkov. V telekomunikacijskem sektorju, ki se trudi storitve prilagajati tako potrebam strank kot novim tehnološkim možnostim, je kompleksnost obračunavanja storitev največja. Prvi so zaznali potrebo, da se sistematično spopadejo z napakami, ki povzročajo izpad dela prihodkov. Oblikovali so celovito ogrodje postopkov, poimenovano 'preprečevanje odtekanja prihodkov'. Po sprostitvi trga z električno energijo kompleksnost storitev distribucije električne energije sledi kompleksnosti telekomunikacijskih storitev. K temu zelo prispeva uvajanje pametnih omrežij s sistemi naprednega merjenja in s pametnimi števci. Velika množica podatkov, ki pri tem nastaja, ponuja možnosti za nove storitve. V magistrski nalogi smo predstavili pristop k preprečevanju odtekanja prihodkov, tako da smo uporabili splošne postopke, pri tem pa upoštevali tudi posebnosti, ki veljajo za dejavnost distribucije električne energije. Posebno pozornost smo namenili vprašanju kakovosti podatkov.

V praktičnem delu naloge smo se osredotočili na pridobivanje znanja iz podatkov, koristnih pri odkrivanju odtekanja prihodkov, še posebej iz podatkov, ki jih s sistemom naprednega merjenja zberemo s pametnih števcev. V podatkovnem skladišču smo te podatke združili s podatki iz obračunskega sistema in iz prostorskega informacijskega sistema. Ob izgradnji skladišča smo odkrili nekaj težav s kakovostjo podatkov, nanje opozorili in nakazali, kako jih odpraviti in kako vzpostaviti mehanizem za spremljanje morebitnih ponovnih pojavov anomalij.

Pri magistrski nalogi smo se osredotočili na pridobitev informacij o značilnostih odjemalcev, nato pa to znanje uporabili pri iskanju morebitnih kraj električne energije. Primerjali smo metode strojnega učenja za razvrščanje dnevnih obremenitvenih diagramov odjemalcev v značilne skupine. Na podlagi analize rezultatov se je kot najboljša izkazala metoda maksimiranja pričakovanj. Hkrati smo določili najprimernejše število skupin z značilno dinamiko dnevne porabe. Po razvrstitvi vseh merilnih mest, za katera smo imeli na voljo izmerjene 15-minutne porabe, smo ugotavljali, katere lastnosti odjemalca najbolj sovpadajo z dinamiko njegove porabe. Ponovno smo preizkusili več metod strojnega učenja in ugotovili, da so za to nalogo najprimernejša odločitvena drevesa. Z uporabo pravil, ki smo jih odkrili, smo ocenili dnevne porabe za vsa preostala merilna mesta. S tako pripravljenimi podatki smo izdelali analitično strukturo, ki je odlična osnova za odkrivanje odtekanja prihodkov. Avtomatizirali smo celoten postopek polnjenja skladišča, odkrivanja in uporabe znanja ter obdelave analitične strukture. Z nekaj primeri aktualnih poročil smo dokazali koristnost tega početja.

Ključne besede:

Preprečevanje odtekanja prihodkov, kakovost podatkov, napredni sistem merjenja (AMI), pametni števci, značilni dnevni obremenitveni diagram, podatkovno rudarjenje, razvrščanje v skupine, klasifikacija, odkrivanje kraje električne energije.

Abstract

The increasing complexity of services also encourages the complexity and heterogeneity of the systems that providers use for measuring and billing these services. The complexity may result in the occurrence of errors and consequently in a revenue leakage. For the telecom industry that strives to adapt their services not only to the needs of customers but also to new technological opportunities the biggest complexity issue is billing the services. They were the first to recognize the need to develop a systematic way to grapple the problem of errors that cause the revenue leakage. The sector prepared a comprehensive framework of procedures called "revenue assurance". With the liberalization of the electricity market, the services of electricity distribution became as complicated as the telecommunications services. This further significantly enhanced the deployment of smart grids, advanced metering infrastructures and smart meters that, with the abundance of data, give opportunities for new services. This master thesis presents the approach we took to grapple with the revenue assurance. We used general procedures and took into consideration the peculiarities of the electricity distribution domain as well. Particular attention was given to data quality issue.

In the practical part of the thesis, we focused on acquiring knowledge from the data that would benefit us in detecting the revenue leakage, from the smart meters' data collected by advanced metering infrastructure in particular. In the data warehouse, the data was combined with the billing system data and the geographic information system data. While building the data warehouse, we encountered some problems with data quality. After we had pointed out the problems, we indicated how to eliminate them and how to establish a mechanism for monitoring any possible recurrences of errors.

The thesis focused on collecting information on the characteristics of consumers. Once we acquired this knowledge, we used it to look for any thefts of electricity. We made a comparison of machine learning methods for the classification of daily load curves of consumers into typical groups. Based on the analysis of the results obtained, we selected the best method, i.e. the expectation maximization method. At the same time, we determined the best number of clusters with the typical dynamics of daily consumption. Once all measuring points with the 15-minute consumption data were classified, we were determining the characteristics of a consumer that coincide the most with the dynamics of his electricity consumption. Again, we tested several machine learning methods and established that decision trees are the most appropriate tool for this task. With established behavior, we estimated daily consumption for all other measuring points. Thus prepared data were used to develop an analytical structure that proves to be an excellent base for discovering the revenue leakage. We automated the entire process of filling the warehouse, finding and applying knowledge and, last but not least, processing the analytical structure. We demonstrated the usefulness of this practice with a few examples of actual reports.

Key words:
revenue assurance, data quality, advance metering system, smart meters, load profiles, data mining, clustering, classification, electricity theft detection

1 Uvod

Pogosto slišimo krilatico, da so podatki bogastvo. To ne drži povsem. Če so samo odloženi, so najprej strošek. Če jih je preveč, so lahko tudi ovira. Vendar podatki so povezani z bogastvom - skrito je v njih. Odkrivati ga pričnemo z raziskovanjem odnosov med podatki. Tako pridemo do informacij. Sedaj lahko odgovorimo na vprašanja: Kdo? Kaj? Kje? Kdaj? A bogastvo se še poveča, ko spoznamo in razumemo vzorce v podatkih in odnosih. Takrat imamo znanje in lahko odgovorimo na vprašanje: Zakaj? A tukaj se še ne konča. Če nam uspe razumevanje še povečati, pridobimo modrost in spoznamo pravila, katerim sledi znanje. Tako nam morda uspe celo napovedati prihodnost.

V magistrski nalogi želimo stopati po opisani poti odkrivanja vrednega v podatkih in združiti znanstveno raziskovalno delo s poslovno uporabno implementacijo.

1.1 Predstavitev problema

S pristopom zagotavljanja prihodka (angl. Revenue Assurance), ki se je uveljavil predvsem v telekomunikacijski branži, želimo raziskati in predstaviti možnosti na tem področju v elektrodistribucijskem podjetju, saj ta pristop v elektroenergetskih vodah vsaj v Sloveniji še ni prisoten v večji meri. Skušali bomo identificirati področja odtekanja prihodkov v raznih poslovnih procesih podjetja in s povezovanjem različnih podatkovnih virov vzpostaviti zametek analitskega sistema, ki bo nudil podporo za preprečevanje tega odtekanja. Pri tem bo pomembno zagotavljanje kakovosti podatkov [1]. Eden od pomembnih virov podatkov, na katerega se bomo osredotočili, bodo odčitki porabljene električne energije, pridobljeni s pametnimi števci [13]. Podjetja za distribucijo električne energije z znatnim investiranjem že več kot 10 let postopoma vgrajujejo pametne števce, s katerimi daljinsko odčitavajo izmerjeno porabo električne energije pri odjemalcih. Števci zajemajo vrednosti meritev vsakih 15 minut, prenos podatkov v osrednjo podatkovno bazo posameznega elektrodistribucijskega podjetja pa se izvede tipično enkrat dnevno. Elektro Celje d. d. z električno energijo oskrbuje slabo petino (18,5 %) odjemnih mest v Sloveniji. Konec leta 2015 so od vseh njihovih aktivnih 170.000 odjemnih merilnih mest, daljinsko odbirali 96.000, poleg tega pa še 2.785 merilnih mest v transformatorskih postajah. Zbrane podatke uporabijo za mesečni obračun porabljene energije. Pri odkrivanju izgub v omrežju, kamor sodijo tudi kraje električne energije, trenutno uporabljajo ad hoc poizvedbe ter znanje in sklepanje ljudi, ki se s tem ukvarjajo.

Velik obseg zbranih podatkov ponuja priložnost za podrobno spoznavanje narave odjemalcev v obliki dnevnega diagrama porabe električne energije. Glede na obliko dnevnega diagrama lahko odjemalce razvrstimo v več skupin z značilnimi dnevnimi diagrami. Doslej so raziskovalci na tem področju v Sloveniji [3, 6] in v tujini [11, 13] določali značilne diagrame

predvsem za večje poslovne uporabnike in to na podlagi relativno majhne množice vzorcev. Zaslediti pa je tudi raziskovalno delo, usmerjeno v obdelavo velikih količin tovrstnih podatkov [7].

1.2 Namen in cilji naloge

V nalogi želimo iz relativno velike množice podatkov odkriti značilne vzorce v izmerjeni porabi električne energije pri odjemalcih in vzorce v proizvodnji električne energije pri razpršenih virih, priključenih na distribucijsko omrežje. Odkriti želimo tudi informacijo o spreminjanju porabe skozi čas (za delovni dan, vikend, praznik, po mesecih, med sezonami in skozi leta). Z odkrivanjem korelacij med podatki želimo raziskati vpliv lastnosti odjemalcev (kot so: vrsta odjemalca – gospodinjski / poslovni, mestni / podeželski, lokacija oz. geografsko področje, priključna moč, dvotarifno obračunavanje, …) na velikost in obliko porabe električne energije. Na ta način iz podatkov odkrito znanje je lahko dobra podlaga tudi za odkrivanje kraj električne energije [8, 14, 15].

Dodatno priložnost za odkrivanje kraj, v kombinaciji s podatki o izmerjenih porabah pri odjemalcih, ponujajo še podatki, izmerjeni na izvodih iz transformatorskih postaj. Ti predstavljajo vsoto energije, ki jo porabljajo vsi odjemalci, priključeni na posamezen izvod ter tehnične izgube v tem delu omrežja.

Podatke bomo morali za analizo pripraviti tako, da bomo izločil preveč pomanjkljive in neregularne, nadomestili manjkajoče, kjer je to smiselno in normirali vrednosti. Te podatke (diagrame s 96 odbirki za vsakih 15 minut v dnevu, za vsakega od odjemalcev z daljinskim odbiranjem števca), bomo skušali razvrstiti po podobnosti v naravne skupine. Ko bomo za vsa odjemna mesta, ki imajo nameščen števec z daljinskim odbiranjem identificirali, kateri tipični skupini pripadajo, bomo skušali odkriti korelacijo med lastnostmi odjemnega mesta ter njegovo obliko in velikostjo porabe energije. Na podlagi teh odkritih pravil bomo v tipične skupine razvrstil tudi vsa odjemna mesta, ki še niso daljinsko odbirana in zanje izračunali njihove nadomestne 15-minutne porabe. Tako bomo zagotovili podatke o porabi za vsa odjemna mesta v omrežju – za tiste z daljinskim odbiranjem dejansko izmerjene, za preostale pa izračunane nadomestne. Iz teh podatkov lahko pripravimo zelo podrobne in točne vhodne podatke sistemu za analizo omrežja z izračunom pretokov moči in padcev napetosti, s katerim na podlagi podrobnega in relevantnega modela, z minimizacijo tehničnih izgub, določamo optimalno obratovalno stanje sistema in tako omejimo eno od pomembnejših odtekanj prihodkov.

Glavni prispevek magistrskega dela bo, da bomo ob uporabi domenskega znanja za področje distribucije električne energije, proučil niz postopkov obdelave in analize podatkov, s katerimi bomo iz velike množice podatkov izluščili uporabne informacije, na podlagi katerih bomo predlagali ukrepe in pristope za izboljšanje poslovnih praks.

2 Preprečevanje odtekanja prihodkov

2.1 Splošno o pristopu 'preprečevanje odtekanja prihodkov'

Izraz 'Revenue Assurance' (RA) iz Angleščine lahko prevedemo kot 'zagotavljanje prihodkov', a vsebinsko ustrezneje je 'preprečevanje odtekanja prihodkov'. Ta izraz se je uveljavil šele po letu 2000, sprva na področju poslovanja telekomunikacijskih operaterjev, ki so pristop razvili pod okriljem združenja TM Forum (TeleManagement Forum) [19]. Znotraj istega združenja je bil razvit tudi okvir poslovnih procesov za udeležence na telekomunikacijskem področju eTOM, ki je postal de-facto standard, ki omogoča prenekatero sodobno komunikacijsko storitev.

RA so definirali kot skupek sistematično zastavljenih aktivnosti za izboljševanje poslovnih procesov in zagotavljanje kakovosti podatkov z namenom zagotavljanja prihodkov. Pri tem se ne osredotočajo na povečanje povpraševanja in povečanega obsega poslovanja, temveč na to, da organizacija opravljene storitve, tudi v celoti pravilno, točno in pravočasno zaračuna ter zagotovi, da jih dobi plačane. [1]

Nasprotno pod izrazom 'odtekanje prihodkov' razumemo, da organizacija ne zaračuna pravilno ali sploh ne zaračuna in/ali ne izterja zasluženega plačila za storitve, ki so bile opravljene. Razlogi za odtekanje prihodkov tičijo tako v težavah s kakovostjo podatkov (kot je npr. nekonsistentnost povezav ali izguba podatkov), kot tudi v težavah s slabo določenimi in izvajanimi poslovnimi procesi. V idealnih razmerah aktivnosti za preprečevanje odtekanja prihodkov spremljajo vsak korak v procesu, s katerim ustvarjamo prihodek. Te aktivnosti so v veliki meri sorodne upravljanju tveganj v podjetju, saj morajo zajemati vsa tveganja, povezana s prihodki.

Izvajanje tehnik zagotavljanja prihodkov (RA) je pomembno ne le zaradi odkrivanja nezaračunanih ali napačno zaračunanih storitev, ampak tudi zaradi razumevanja in nenazadnje odprave razlogov za takšne neželene pojave.

V [20] so bili določeni različni pristopi k RA:

- Reaktivni pristop k RA; uporabljamo ga le za odkrivanje obstoječih odtekanj prihodkov. Aktivnosti so usmerjene v identifikacijo in odpravo vzrokov za dejanske izgube prihodkov, ki so se že zgodile.
- Aktivni pristop k RA; naslavlja nepravilnosti takoj, ko se zgodijo in proži korektivne aktivnosti z namenom, da prepreči kakršne koli izgube. Poslovni proces med izvajanjem sproti nadziramo. Sprotno odkrivanje težav in takojšnja korekcija posledičnih napak skuša preprečiti odtekanje prihodkov, preden te povzročijo škodo in prizadenejo odjemalca.
- Proaktivni pristop k RA; deluje na podlagi prediktivne analitike. Z uvedenimi kontrolnimi merili skuša vnaprej preprečiti nastanek težav.

Vsi trije pristopi so komplementarni - se dopolnjujejo. Pomembno je, da v prvem koraku zaznamo in odpravimo dejanska odtekanja prihodkov v podjetju. Ko odkrijemo razloge za to, z uvedbo aktivnega ali proaktivnega pristopa RA preprečimo nadaljnjo škodo in ciljno vnaprej preprečimo nastanek le te.

Kako preprečujemo odtekanje prihodkov?

- Za vsako aktivno storitev identificiramo ključne atribute, ki vplivajo na prihodek.
- Za vsak ključni atribut identificiramo vir podatkov.
- Povežemo vire podatkov v informacijski model.
- Identificiramo tok podatkov od virov do ciljev (ponorov).
- Določimo kontrole, ki bodo zagotovile integriteto podatkov čez celoten tok.
- Določimo mejne vrednosti toleriranih izgub v vsaki točki, ki izhajajo iz ciljne skupne mejne vrednosti.
- Zasnujemo mehanizem navzkrižnih kontrol vzdolž celotnega podatkovnega toka.

2.1.1 Stopnje zrelosti preprečevanja odtekanja prihodkov

TM forum za podjetja (za telekomunikacijske operaterje) določa stopnje zrelosti pri preprečevanju odtekanja prihodkov v treh vidikih: strateškem, organizacijskem in izvedbenem.

Po strateški plati je podjetje na najnižji, prvi ravni, če glede odtekanja prihodkov nima opredeljene strategije. Stopnja zrelosti narašča preko vse bolj formaliziranih strategij in uveljavljenega vpliva tudi na nižjih vodstvenih ravneh, do najvišje 5. stopnje pri strategiji, ki je celovito formalizirana, temelji na tveganjih in vključuje parametre zmanjševanja stroškov.

Nizka stopnja zrelosti po organizacijski plati pomeni, da je področje preprečevanja odtekanja prihodkov prepuščeno posameznikom, pa še ti se ne ukvarjajo izključno s tem. Visoka stopnja organizacijske zrelosti postavlja RA v nadzorno in svetovalno vlogo. Zaposleni v namenskem oddelku imajo tudi računovodska in revizorska znanja.

Izvedbeni vidik vključuje procese in orodja. Nizka stopnja zrelosti pomeni, da se s preprečevanjem odlivanja prihodkov ukvarjajo šele, ko se je to že zgodilo in so ga zaznali (reaktivni pristop), pri tem pa uporabljajo znatne 'ročne' napore le s programsko opremo pri končnih uporabnikih. Višanje stopnje zrelosti v tem vidiku se odraža v vse bolj celovitih in formaliziranih procesih, v katerih so aktivnosti v veliki meri optimizirane in avtomatizirane na osrednji analitični platformi.

2.1.2 Kakovost podatkov

Kakovost podatkov (angl. data quality) določajo lastnosti, kot so točnost (angl. accuracy), pravilnost (angl. correctness), celovitost (angl. completeness), pravočasnost (angl. timeliness and currency), ustreznost (angl. relevance) in konsistentnost (angl. consistency) [24]. V obdobju naraščajočih količin podatkov in vse kompleksnejših procesov, je obvladovanje kakovosti podatkov vse pomembnejše in vse zahtevnejše. Pri odkrivanju odtekanja prihodkov moramo obravnavati celoten procesni tok in ob tem podatkovni tok, vse od zajema vhodnih

podatkov (meritev, priključitev) do izstavljanja računa. Proučiti moramo, ali obstajajo storitve in izdelki, ki so bili realizirani, a niso bili pravilno ali v celoti ali sploh ne zaračunani stranki. Ločimo dve vrsti napačnega zaračunavanja:

- 'podzaračunavanje': odjemalec naroči in prejme določen produkt, vendar ne dobi računa zanj. V nekaterih primerih odjemalec prejme in plača račun za drug, cenejši produkt.
- 'prezaračunavanje': predstavlja primere, ko odjemalec prejme neupravičeno previsok račun, na katerem so postavke, katerih ni prejel ali pa so tiste, ki jih je prejel, obračunane z napačnimi, višjimi tarifami. V takšnih primerih se stranke veliko pogosteje pritožijo, kot pri podzaračunavanju.

2.1.3 Čiščenje podatkov

Odkrivanje težav s kakovostjo podatkov samo po sebi ne zadošča. Po identifikaciji narave in obsega težav moramo opraviti čiščenje. Ločimo med dvema vrstama težav:

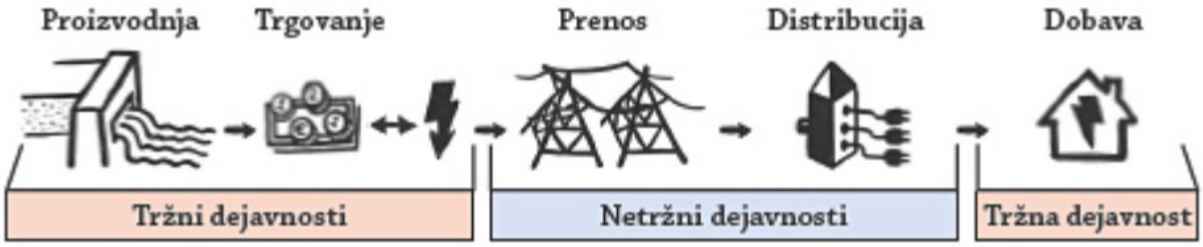
- Zgodovinske težave so nekonsistentnosti, ki izvirajo iz migracij, uvedb novih sistemov ali procesov ter iz uvedbe procesnih sprememb. Ko odpravimo nekonsistentnosti zgodovinskih težav, se te ne bodo pojavile ponovno.
- Procesne težave so težavnejše. Nekonsistentnosti v podatkih izvirajo iz nepravilnosti v procesu in se pojavljajo znova in znova. Ključnega pomena za odpravo težav je odkriti in izločiti izvorni vzrok zanje. Ko popravimo proces (sistem), lahko težavo obravnavamo kot zgodovinsko in pristopimo k čiščenju podatkov samih.

Ne zadošča, če obravnavamo vsak vidik kakovosti podatkov posebej. Učinkovita obravnava upošteva možnost, da je napaka na odjemalčevem računu posledica kombinacije več tipov težav s podatki hkrati.

2.2 Preprečevanje odtekanja prihodkov v elektrodistribuciji

2.2.1 Organiziranost trga z električno energijo

V verigi, ki odjemalcu zagotovi električno energijo, nastopa več deležnikov.



Slika 2.1: Shematski prikaz udeležencev na trgu z električno energijo (vir: http://www.agen-rs.si/udelezenci-na-trgu-z-elektricno-energijo)

Za proizvodnjo električne energije skrbijo proizvajalci v elektrarnah z uporabo različnih virov – obnovljivih (npr. voda, veter, sonce) in neobnovljivih (npr. premog, nafta, plin, jedrsko gorivo).

Prenos večjih količin energije na daljše razdalje izvajamo s prenosnim omrežjem, na višjih napetostnih nivojih. V Sloveniji opravlja naloge sistemskega operaterja visokonapetostnega prenosnega omrežja (400, 220 in 110 kV) javno podjetje ELES, d. o. o. .

Z elektrodistribucijskim sistemom pretvorimo električno energijo na nižje napetostne nivoje in jo prenesemo do končnih odjemalcev. Distribucijska omrežja obsegajo elektroenergetske vode in naprave na nizko napetostnem nivoju (0,4 kV), srednje napetostnem nivoju (10, 20 in 35 kV) in v posameznih primerih tudi na visoko napetostnem nivoju (110 kV). Lastniki teh sistemov v Sloveniji, so elektrodistribucijska podjetja (EDP):

- Elektro Celje, d. d.
- Elektro Gorenjska, d. d.
- Elektro Ljubljana, d. d.
- Elektro Maribor, d. d.
- Elektro Primorska, d. d.

Z odprtjem trga električne energije v Sloveniji je bilo ustanovljeno podjetje SODO, d. o. o. - sistemski operater distribucijskega omrežja, kateremu je bila podeljena koncesija za upravljanje z distribucijskimi omrežji. Na podlagi pogodbe o najemu infrastrukture za distribucijo električne energije SODO, d. o. o. najema infrastrukturo od elektrodistribucijskih podjetij. Po tej pogodbi EDP zagotavljajo storitve na distribucijskih omrežjih in izvajajo vzdrževanje, razvoj, izgradnjo ter vodenje in obratovanje. EDP izvajajo tudi merjenje energije, katero predajo odjemalcem in na podlagi izmerjenih količin obračunajo omrežnino. Odprt trg električne energije pomeni, da odjemalci lahko prosto izbirajo svojega dobavitelja. Dobavitelji so podjetja, ki na eni strani kupijo energijo od proizvajalcev, na drugi pa jo prodajo odjemalcem in jo na podlagi podatkov o količinah in omrežnini, katere prejmejo od EDP, zaračunajo.

Končni znesek za plačilo dobavljene električne energije za odjemalca je sestavljen iz naslednjih postavk:

- **cene električne energije**, katero določa dobavitelj električne energije in se oblikuje prosto na trgu. Obračuna se za vsako porabljeno kilovatno uro in to odvisno od uporabnikove izbire lahko po enotni (ET) tarifi ali dvotarifno, po v večji (VT) in manjši (MT) tarifi.
- **omrežnine**: Odjemalec električne energije plača za prenos in distribucijo električne energije po električnem omrežju do njegovega prevzemno-predajnega mesta. Omrežnina je namenjena izvajanju dejavnosti distribucijskega operaterja (SODO, d. o. o.), dejavnosti sistemskega operaterja (ELES, d. o. o.), pokrivanju stroškov sistemskih storitev (ELES, d. o. o.) ter pokrivanju stroškov delovanja Agencije za energijo.
  Omrežnina je sestavljena iz:
  - cene za obračunsko moč v kilovatih, ki je odvisna od moči vgrajenih varovalk, in

  - cene za omrežnino, ki se obračuna za vsako porabljeno kilovatno uro električne energije.
- **prispevkov:**
  - prispevek namenjen spodbujanju proizvodnje električne energije iz obnovljivih virov in soproizvodnje z visokim izkoristkom. Obračuna se glede na obračunsko moč.
  - prispevek namenjen povečevanju energetske učinkovitosti. Obračuna se za vsako porabljeno kilovatno uro električne energije.
  - prispevek za delovanje operaterja trga - Borzen, d. o. o.
    Prispevke določa Vlada RS.
- **trošarine** na električno energijo in
- **davka** na dodano vrednost.

Odjemalec ima, ne glede na izbranega dobavitelja, pravico do plačila storitev dobave električne energije in uporabe omrežja z enotnim računom, na katerem uporabo omrežja zaračuna dobavitelj v imenu in za račun operaterja sistema. Na enotnem računu so prav tako zavedene postavke iz naslova plačila prispevkov in trošarine.

Za elektrodistribucijsko podjetje je ključen vir prihodkov omrežnina oz. tisti del le te, ki je namenjen izvajanju dejavnosti distribucijskega operaterja (SODO). Ta namreč na podlagi pogodbe o najemu infrastrukture določen delež te omrežnine nameni za upravljanje, vzdrževanje in nadaljnji razvoj te infrastrukture. 'Pravila igre' za to delitev regulira Agencija RS za energijo, vsakič za triletno obdobje, z Aktom o metodologiji za določitev regulativnega okvira in metodologiji za obračunavanje omrežnine za elektrooperaterje [21]. Metodologiji sta določeni na način, da spodbujata učinkovitost elektrooperaterjev in učinkovitost uporabe sistema.

Metoda reguliranega letnega prihodka in reguliranih omrežnin se izvaja tako, da se za regulativno obdobje elektrooperaterju določi regulativni okvir tako, da omrežnina skupaj z drugimi prihodki iz opravljanja dejavnosti elektrooperaterja in upoštevaje ugotovljeni kumulativni presežek oziroma primanjkljaj omrežnin elektrooperaterja iz preteklih let pokrije načrtovane **upravičene stroške** elektrooperaterja, ob upoštevanju vseh predvidenih okoliščin stroškovno učinkovitega poslovanja elektrooperaterja.

2.2.2 Procesi, povezani z obračunavanjem omrežnine

Kot smo videli v prejšnji točki, je glavni vir prihodkov elektrodistribucijskega podjetja omrežnina, ki se obračunava sorazmerno s količino predane in prevzete energije in s priključno močjo. Zaradi tega se bomo na kratko seznanili s procesi, neposredno povezanimi z obračunavanjem omrežnine, saj se bomo v nadaljevanju naloge osredotočili predvsem na s tem povezana odtekanja prihodkov.

Postopek priključitve na distribucijsko omrežje

V skladu z zakonom o graditvi objektov se gradnja zahtevnega ali manj zahtevnega objekta lahko začne na podlagi pravnomočnega gradbenega dovoljenja.

Za pridobitev gradbenega dovoljenja za objekt, ki se bo gradil na območju, ki se ureja po prostorskih aktih, mora investitor (uporabnik) od elektrodistribucijskega podjetja pridobiti:

- projektne pogoje (za umestitev v prostor, če bo objekt stal v varovalnem pasu gospodarske javne elektro infrastrukture in za priključitev),
- soglasje na projektne rešitve.

Po pridobitvi dokončnega gradbenega dovoljenja mora uporabnik urediti oz. zaprositi za:

- Soglasje za priključitev, v katerem so natančno določeni tehnični pogoji in parametri priklopa (kot npr. priključna moč).
- Pogodbo o priključitvi, s katero se urejajo medsebojna razmerja v zvezi s premoženjskimi vprašanji v zvezi s priključkom, vzdrževanjem priključka in druga medsebojna razmerja, ki zadevajo priključek in priključitev. Ob pogodbi o priključitvi uporabnik plača omrežnino za priključno moč (ki je enkraten strošek, namenjen razvoju omrežja) ter neposredne stroške priključevanja.
- Nadzor nad izdelavo priključka s strani elektrodistribucijskega podjetja, ki vključuje spremljanje gradnje priključka v skladu s predpisi za gradnjo tovrstnih objektov, izvajanje potrebnih stikalnih manipulacij, obveščanje prizadetih kupcev o morebitnih motnjah dobave električne energije, izvajanje priključitve ter druga dela v zvezi z izgradnjo priključka in priključitvijo.
- Sklenitev pogodbe o dobavi ali prodaji električne energije, katero lahko z dobaviteljem električne energije sklene imetnik soglasja za priključitev ali tretja oseba, ki ji je imetnik soglasja za priključitev izdal soglasje za sklenitev pogodbe o dobavi za to prevzemno-predajno mesto. Lastnik proizvodne naprave mora imeti sklenjeno pogodbo o prodaji proizvedene električne energije.
- Sklenitev pogodbe o dostopu do distribucijskega omrežja - elektrodistribucijsko podjetje jo sklene z imetnikom soglasja za vsako prevzemno-predajno mesto potem, ko ima z dobaviteljem urejeno pogodbo o dobavi električne energije in preden izvede priključitev. Ta pogodba določa pogoje dostopa do omrežja, način izvajanja meritev ter zaračunavanje in plačevanje uporabe omrežja in prispevkov. Pogodbo o dostopu v uporabnikovem imenu z distributerjem največkrat sklene dobavitelj.
- Tehnični pregled priključka in merilnega mesta, kjer se preveri skladnost zahtev iz soglasja za priključitev in predhodno izpolnjene vloge za priključitev in dostop do distribucijskega omrežja.
- Priključitev merilnega mesta na omrežje, pri čemer se parametrira in preizkusi delovanje števca in komunikacijske naprave. Hkrati se skupaj z uporabnikom oceni višina bodoče mesečne porabe električne energije, ki se jo zaračunava v akontacijah do prvega rednega odčitavanja števca za obračun električne energije.

Po vseh naštetih korakih, ki jih uporabnik opravi enkrat ob priključitvi (ali nekatere od njih ponovno ob spremembi parametrov pogodbe o dostopu), se prične odjem energije in ciklični proces merjenja porabe, obračunavanja omrežnine in spremljanja plačil ter po potrebi izvajanja izterjave.

Odjemalcem, ki imajo na odjemnem mestu nameščen pametni števec z daljinskim odčitavanjem, vključenim v napredni sistem merjenja, se omrežnina obračunava po dejanski porabi. Ostalim odjemalcem z 'ročnim' odčitavanjem popišejo števčna stanja enkrat letno (ali pogosteje, če pride do spremembe pogodbe o dostopu). Takrat se izračuna povprečna poraba, po kateri se obračunava akontacija vnaprej, hkrati pa se naredi poračun med porabljenimi in že plačanimi količinami (akontacijami). Odjemalci s klasičnimi indukcijskimi števci, lahko tudi sami odčitajo stanje in ga javijo. V tem primeru se tudi njim lahko obračunava po dejanski porabi. V vsakem primeru pa vsaj enkrat letno odčitavanje opravijo delavci elektrodistribucijskega podjetja.

Po obdelavi obračuna porabljenih količin energije in s tem povezano omrežnino in prispevki, distributer podatke v predpisani standardni obliki, (ti. 'prilogi A' https://www.sodo.si), preko enotne vstopne točke za izmenjavo podatkov med udeleženci na trgu električne energije, posreduje dobaviteljem (vsakemu za 'njegove' odjemalce). S temi vhodnimi podatki dobavitelji izdelajo dodatni obračun, kjer za količine, ki jih je distributer izmeril in posredoval, skladno s sklenjenimi pogodbami o dobavi, obračunajo še samo energijo. Večina uporabnikov pooblasti dobavitelja, da jih zastopa pri izvajalcu distribucije v postopkih dostopa do omrežja. V tem primeru prejme le po en račun za obdobje in sicer dobaviteljevega, ki pa vsebuje združene postavke energije, omrežnine, prispevkov in trošarine. Ostali uporabniki prejemajo ločen račun za energijo od dobavitelja in ločenega za omrežnino od distributerja. Spremljanje plačil in proženje procesa izterjave izvaja izdajatelj računa.

2.2.3 Načini odtekanja prihodkov v elektrodistribuciji

Možnosti za odtekanje prihodkov v elektrodistribuciji je tako kot v vsakem poslu, mnogo. Kot že omenjeno v 2.1., je področje preprečevanja odtekanja prihodkov povezano z obvladovanjem tveganj neuresničitve poslovnega izida. Kjer koli nekega tveganja z ustreznimi ukrepi ne uspemo obvladati, se lahko dogodi odtekanje prihodkov.

Pri izvajanju procesov, opisanih v prejšnjih točkah, lahko povzročajo odtekanje prihodkov različne anomalije, v kolikor se pojavijo.

Možne nepravilnosti v procesih:

- Zamuda pri vnosu podatkov o pogodbenem razmerju v obračunski sistem (npr. odjemno mesto je priključeno, pogodba o uporabi sistema še ni aktivna), oz. nespoštovanje pravil (npr. opravljena priključitev brez veljavne pogodbe o dostopu).
- Nov števec je nameščen (meritve se izvajajo) a v obračunskem sistemu (eIS) ni evidentirana zamenjava števca.
- Števec, ki je daljinsko čitan, ima vpisano napačno tovarniško številko (to se lahko dogaja le pri sistemu, ki zahteva ročni vnos tovarniških številk in ne pri tistem, kjer se števci samodejno registrirajo in sporočijo svojo tovarniško številko).

Možna slaba kakovost podatkov:

- Pokvarjen števec (npr. ne meri posamezne tarife - stanje le te se ne spreminja, …)

- Napake pri evidentiranju tovarniške številke števcev, vgrajenih v transformatorskih postajah ter nepravilno evidentiranje mesta umestitve v omrežju.
- Anomalije v opisu topologije omrežja (BTP), kar povzroča odstopanja pri seštevanju porabe izmerjene pri odjemalcih in primerjanju te sumarne količine z izmerjeno v transformatorski postaji.
- Napake v podatkih o mestu priključitve merilnega mesta v omrežje.
- Napačna definicija odbirka v eIS (vpisano napačno število mest, od tega decimalnih mest…, neustrezno prestavno razmerje pri merjenju energije in pri merjenju moči).
- Nepravilnosti v pogodbi o uporabi sistema z odjemalcem (npr. nepravilen status pogodbe, neustrezen datum pričetka ali konca veljavnosti pogodbe).
- …

Goljufije, kraje:

- Stranka ima vgrajene večje varovalke, kot jih dopušča obračunska moč (pri vsakem obračunu bi morala plačati večje nadomestilo za priključno moč).
- Stranka ima vgrajene večje varovalke, kot je priključna moč (zmanjšati mora odjem in varovalke ali pa pridobiti soglasje in pogodbo o priključitvi za večjo priključno moč, plačati omrežnino za priključno moč in nato spremeniti pogodbo o dostopu - povečati obračunsko moč – višje bo tudi mesečno plačilo).
- Stranka, kateri merimo konično moč, presega zakupljeno konično moč.
- …

Tehnične izgube:

Na elektrodistribucijski sistem lahko gledamo kot na zaključen sistem, v katerega vhod predstavlja pritekajoča energija (od prenosnega omrežja in od proizvajalcev električne energije, priključenih neposredno na distribucijsko omrežje), izhod pa vsota vse energije, oddane odjemalcem. V realnem sistemu v energijski bilanci med izhodom in vhodom vedno obstaja razlika. To so izgube, katere nastajajo zaradi povsem fizikalnih lastnosti omrežja in jih imenujemo '**tehnične izgube'**.

Razliko med količino energije, ki jo odjemalci prejmejo iz distribucijskega sistema in tisto, ki jo uspe distributer zaračunati in jo dobi tudi dejansko plačano, imenujemo '**komercialne izgube'**. Razlog zanje so anomalije, opisane v prejšnjih točkah.

Tudi na del tehničnih izgub lahko na nek način gledamo kot na odtekanje prihodkov, saj jih lahko do določene mere zmanjšamo z optimizacijo električnih lastnosti omrežja. Pomembno vlogo pri optimizaciji ima informacijski sistem za upravljanje distribucijskega sistema (DMS – Distribution Management System), s katerim izdelamo podroben model električnega sistema in z analizami pretokov moči, padcev napetosti, kratkostičnih tokov in podobnimi, skušamo doseči optimalna razklopna stanja omrežja. Hkrati pri načrtovanju dograditev in sprememb v omrežju le te načrtujemo tako, da so celotne tehnične izgube čim manjše. Vendar minimalne izgube niso edini optimizacijski kriterij, temveč upoštevamo tudi čim višjo zanesljivost in

razpoložljivost ter ob izvajanju preklopov tudi skušamo najti takšno zaporedje aktivnosti, da bo število odjemalcev, ki bodo zaznali motnjo, čim manjše oz., da bo izpadla energija čim manjša.

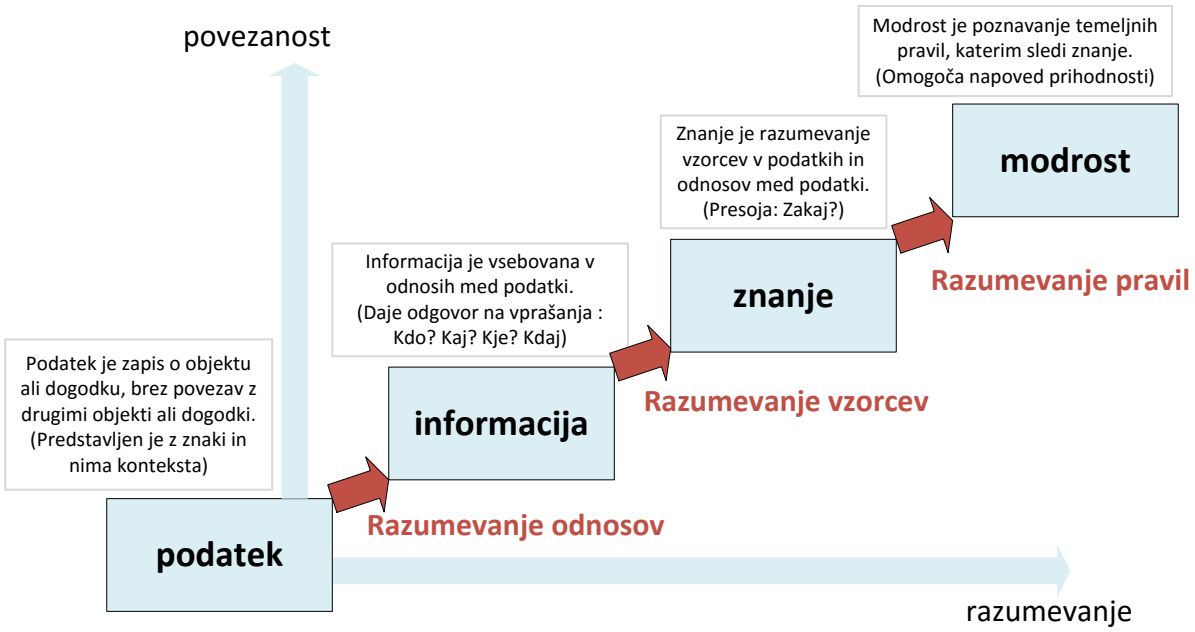
V nalogi se zmanjševanja tehničnih izgub ne bomo dotikali, saj bi celovita obravnava zahtevala veliko povečanje obsega, ki pa tudi sicer ne sovpada s predmetom naloge. Omenimo le, da je pomembno, da sistemu DMS zagotovimo kakovostne vhodne podatke o obremenitvah v posameznih točkah omrežja in o značilnostih bremen (odjemalcev). Te podatke lahko zelo dobro priskrbimo iz števčnih meritev, ki jih izvajamo pri odjemalcih in v transformatorskih postajah v omrežju, kar pa je obravnavano v nalogi.

2.2.4 Predlagan model za preprečevanje odtekanja prihodkov

Elektrodistribucijska podjetja že ves čas izvajajo določene aktivnosti in imajo vzpostavljene procese, s katerimi odkrivajo in preprečujejo morebitne napake v obračunavanju količin električne energije in s tem povezane omrežnine. Prav tako v okviru zakonsko predpisanih možnosti izvajajo postopke izterjave neplačanih dolgov. Vendar pristop preprečevanja odtekanja prihodkov, kot so ga pod okriljem TM foruma zastavili ponudniki telekomunikacijskih storitev [19], ponuja elektrodistributerjem v razmislek kar nekaj možnih izboljšav v smislu celovitejšega zavedanja in uvedbe vseh pristopov: reaktivnega, aktivnega in proaktivnega.
Dodatno priložnost ponujajo vse bolj številni podatki, katere ob zagotavljanju storitev zberejo. Tukaj imamo v mislih predvsem podatke o števčnih meritvah iz sistema naprednega merjenja (AMI). Seveda pa 'poplava podatkov' sama po sebi še ne pomeni nujno izboljšanja. Ključno za sprejemanje ustreznih odločitev je, da iz podatkov in relacij med podatki izluščimo čim več informacij ter iz teh razvijemo znanje o vzorcih, ki se pojavljajo v podatkih in pravilih, katerim sledijo. To je lahko podlaga za odločanje, tudi avtomatsko. Opredelitev konceptov podatek, informacija, znanje, modrost ter razmerja med njimi, prikazuje slika 2.2.

V nadaljevanju naloge se bomo bolj kot na procese odkrivanja in preprečevanja odtekanja prihodkov in s tem povezane delovne tokove, osredotočili na podatkovno plat zgodbe.
V podatkovno skladišče bomo zbrali in medsebojno povezali podatke o osnovnih konceptih iz več sistemov, ki sodelujejo v procesu obračunavanja omrežnine. Ob izgradnji skladišča bomo skušali identificirati nepravilnosti v podatkih, opozoriti nanje, jih popraviti ali izločiti ter odkriti vzroke za njihov nastanek.
Nad podatkovnim skladiščem bomo zgradili analitsko strukturo (angl. OLAP - Online analytical processing), s katero bo omogočeno poročanje in izvajanje analiz nad podatki daljšega časovnega obdobja.
Podatke o izmerjenih količinah energije bomo preoblikovali v informacijo o dnevni dinamiki porabe energije na merilnih mestih. Z upoštevanjem te informacije in metod strojnega učenja, bomo ugotovili, kakšne so tipične oblike dnevne dinamike porabe energije ter skladno z njo v skupine razvrstili vsa merilna mesta.

Za merilna mesta v posamezni skupini bomo skušali najti skupne lastnosti oziroma dejavnike, ki najbolj vplivajo na dinamiko porabe. S tem znanjem bomo bolje razumeli obnašanje odjemalcev. Zasledovali bomo tudi trende spreminjanja teh navad skozi čas.
V nalogi bomo nekatere od teh elementov realizirali do stopnje, primerne za redno poslovno uporabo pri odkrivanju in preprečevanju odtekanja prihodkov.



Slika 2.2: Opredelitev konceptov podatek, informacija, znanje, modrost in razmerja med njimi (povzeto po: https://en.wikipedia.org/wiki/DIKW_Pyramid)

2.3 Pametni števci

Sistem naprednega merjenja (angl. AMI – Advanced Metering Infrastructure) s 'pametnimi števci' je eden od osnovnih delov infrastrukture, za katero se je uveljavil izraz 'pametna omrežja'. Izraz, ki je direktni prevod splošno uveljavljenega angleškega izraza 'SmartGrids', predstavlja elektroenergetsko omrežje, ki lahko stroškovno učinkovito vključuje vse proizvodne vire, odjemalce in tiste, ki so oboje, s ciljem ekonomsko učinkovitega trajnostnega sistema z nizkimi izgubami ter visokim nivojem zanesljivosti, kakovosti in varnosti dobave električne energije [17]. Pametno omrežje se lahko hitro odziva na povečano porabo ali na povečano proizvodnjo električne energije.

Vlaganje v pametna omrežja lahko zmanjša potrebe po siceršnjih investicijah za povečanje zmogljivosti omrežja, hkrati pa lahko pripomore k doseganju ambicioznih okoljskih ciljev, katerim je Slovenija zavezana.

Kot 'pametne števce' v elektrodistribuciji smatrajo števce, ki imajo vsaj naslednje lastnosti oz. zmožnosti:

- merjenje skupne delovne energije v vseh fazah, v obeh smereh (A- in A+),

- merjenje skupne jalove energije v vseh fazah, v obeh smereh (R- in R+),
- merjenje trenutnih in povprečnih moči v vseh fazah skupaj,
- merjenje napetosti po fazah,
- merjenje toka po fazah,
- merjenje trenutne frekvence,
- merjenje faktorja moči,
- merjenje prenapetosti, podnapetosti in izpadov napetosti ter beleženje teh dogodkov,
- hranjenje obremenilnega diagrama (s časovno značko, statusom in vrednostmi A+, A-, R+ in R-) z nastavljivo periodo (tipično 15 minut) za obdobje 40 dni,
- omejevanje porabe (vezano na A+),
- daljinski vklop in izklop porabe (preko programabilnih relejnih izhodov),
- beleženje dogodkov, vezanih na odklopnik,
- zaznavanje in beleženje posebnih dogodkov (npr. prisotnost škodljivega magnetnega polja, odprtje pokrova števca ali vstop v omarico – zaznavanje preko programabilnih vhodov),
- možnost nastavitve več časovnih tarif in tarifnih programov,
- daljinsko parametriranje števca,
- daljinsko nalaganje/posodabljanje aplikativne programske opreme (ne metrološke),
- vmesnik za hišne prikazovalnike in ostale naprave hišne avtomatizacije,
- vmesnik za povezavo z ostalimi števci (plin, toplota, voda,…),
- komunikacijski vmesnik za dvosmerno komunikacijo z merilnim centrom (tipično protokol: DLMS, vmesnik: PLC S-FSK).

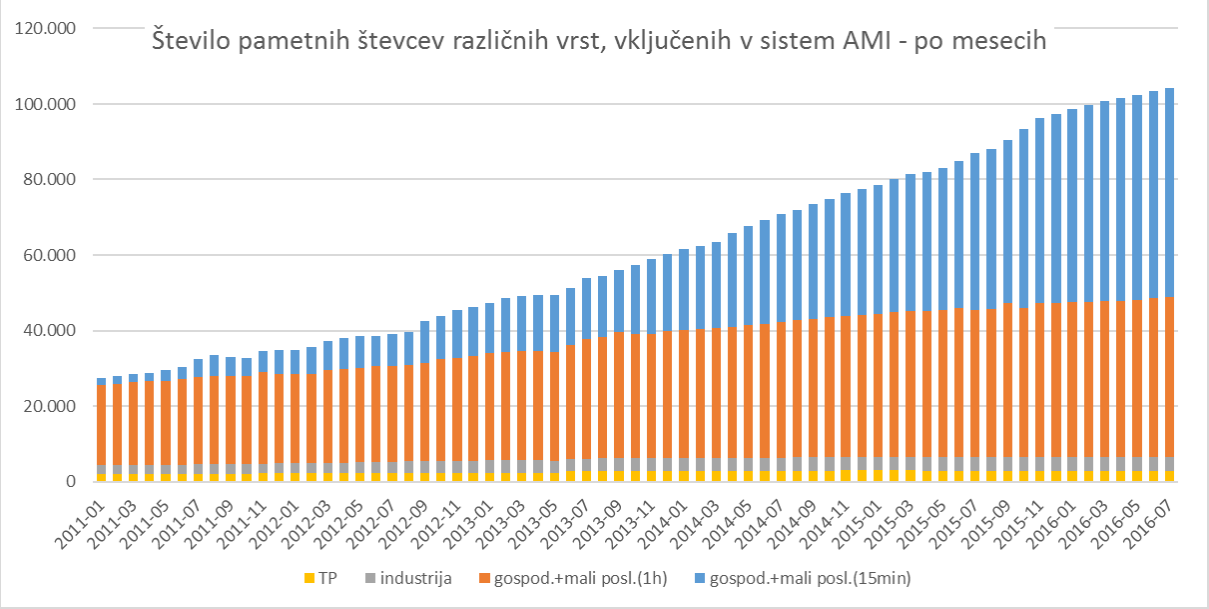
Po direktivi EU 2009/72/ES, mora Slovenija do leta 2020 opremiti z naprednimi števci električne energije vsaj 80% vseh odjemalcev.

Uvedba sistema naprednega merjenja in pospešeno nadomeščanje klasičnih indukcijskih števcev s 'pametnimi' elektronskimi števci za elektrodistribucijsko podjetje predstavlja veliko naložbo. Vendar je po dosedanjih izkušnjah pričakovati relativno hitro povrnitev vloženih sredstev. Koristi od uvedbe pametnih števcev se kažejo kot:

- nižji stroški odčitavanja podatkov in manjša možnost napak,
- boljši pregled nad porabo energije, s čimer lahko veliko učinkoviteje odkrivamo krajo,
- možnost krmiljenja porabe in s tem znižanje koničnih obremenitev (s čimer bi lahko predvidoma zmanjšali ali preložili zahteve po povečevanju zmogljivosti omrežja),
- možnost uvedbe novih paketov za prodajo električne energije pri dobaviteljih, na podlagi izmerjene dejanske porabe,
- nižji stroški merjenja in priključevanja razpršenih virov proizvodnje električne energije,
- možnost spremljanja dodatnih parametrov kakovosti energije,
- …

Dinamiko vključevanja pametnih števcev v sistem naprednega merjenja v Elektru Celje prikazuje slika 2.3. Število vgrajenih pametnih števcev na merilnih mestih je dejansko večje, kot jih prikazuje diagram, vendar nekateri zaradi omejitev pri vzpostavljanju komunikacije, še

niso vključeni v sistem. Trenutno je v omrežju Elektra Celje delež pametnih števcev, za katere se redno izvaja daljinsko odčitavanje, približno 62% vseh števcev. Z nadaljevanjem vključevanja z enakim tempom, bo v Elektru Celje do konca leta 2020 okoli 90% vseh števcev pametnih.



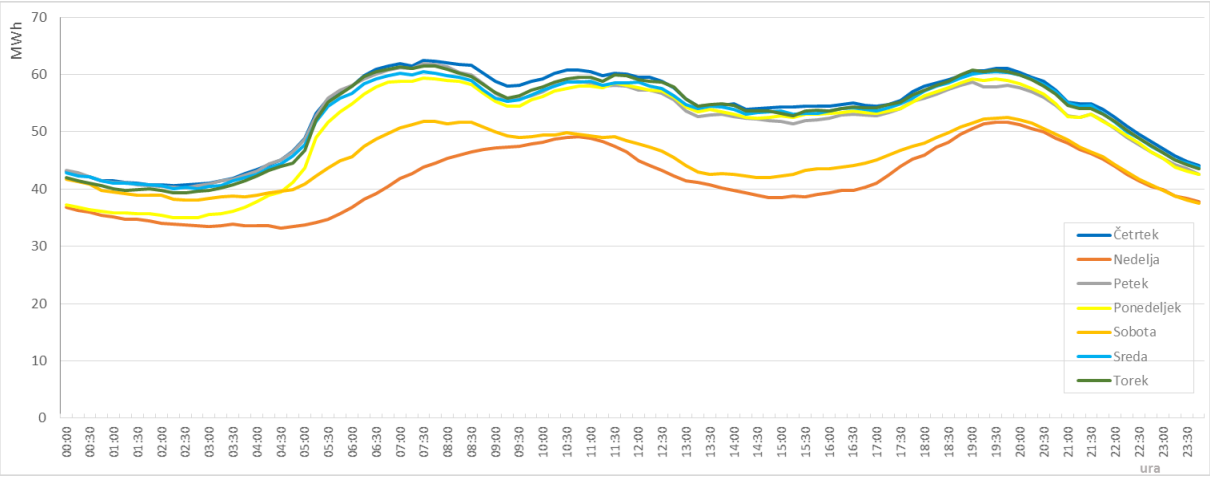
Slika 2.3: Rast števila pametnih števcev različnih vrst, vključenih v sistem AMI Elektra Celje

Sistem naprednega merjenja (AMI) s 'pametnimi števci' tako predstavlja tudi pomemben element, ki nam lahko zelo pomaga pri preprečevanju odtekanja prihodkov.

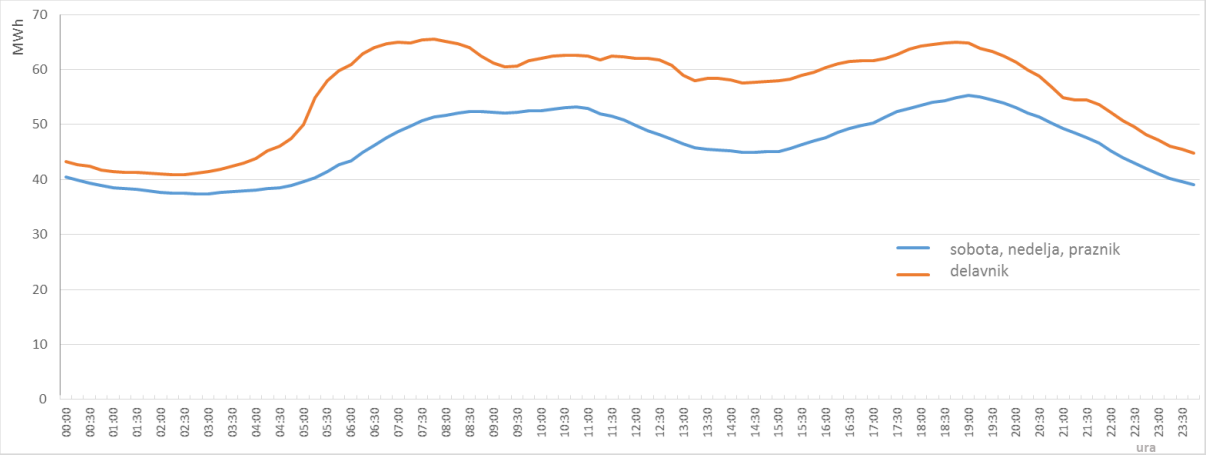
2.4 Dnevni obremenilni diagrami

Dnevni obremenilni diagram je diagram porabe električne energije v enem dnevu. Abscisna os predstavlja čas, ordinatna pa količino električne energije. V diagramih se lahko odločimo za podajanje vrednosti v različnih časovnih intervalih. Najpogostejši so 15, 30 ali 60 minutni intervali. V nalogi smo se odločili za 15-minutni interval, kar pomeni, da ima dnevni obremenilni diagram 96 časovnih točk.

Odločitev za takšen interval je povsem naravna, saj so se elektrodistributerji ob uvedbi nameščanja daljinsko odbiranih 'pametnih' števcev (kar se je pričelo, sprva pri večjih odjemalcih, okrog leta 2002), odločili, da bodo zajemali meritve energije s 15-minutnim intervalom. Tako sedaj pretežno (povsod, kjer števec to omogoča) zajemajo po 96 kompletov meritev na dan. Kot komplet meritev je mišljen set štirih vrednosti: odtekajoče delovne energije A+, pritekajoče delovne energije A-, odtekajoče jalove energije R+, in pritekajoče jalove energije R- (ob teh merijo in beležijo tudi druge veličine in dogodke).



Slika 2.4: Primer povprečnih dnevnih diagramov vse energije, prevzete v omrežje Elektra Celje po posameznih dneh v tednu, za drugi kvartal leta 2016



Slika 2.5: Primer dnevnih diagramov vse energije, prevzete v omrežje Elektra Celje za povprečen delovnik in povprečen dela prost dan v letu 2015

Pred odprtjem trga z električno energijo so se distribucijska podjetja in posamezni odjemalci v Sloveniji z dnevnimi obremenilnimi diagrami in t.i. 'voznimi redi' ukvarjali predvsem v obratovalne namene, na nivoju celotnega omrežja ali določenega segmenta omrežja.

Odpiranje trga z električno energijo se je v Sloveniji pričelo v letu 2000. Elektrodistribucijska podjetja so morala z letom 2001 o izvajanju upravljanja distribucijskega omrežja in o izvajanju tržne dejavnosti (vključujoč prodajo električne energije), voditi ločene računovodske izkaze. Z aprilom 2001 so večji odjemalci (s priključno močjo nad 41 kW), postali upravičeni do proste izbire dobavitelja in do dogovarjanja o ceni električne energije (skozi letne pogodbe). Od leta 2003 lahko t.i. 'upravičeni odjemalci' kupujejo energijo tudi v tujini. Cena električne energije je bila odvisna od napovedanih količin in dinamike porabe ter odstopanj od napovedi. Takrat se je pojavila potreba po 'nadomestnih obremenilnih diagramih', ki predstavljajo način za določitev dnevne krivulje porabe za tiste odjemalce, ki nimajo nameščenega števca, ki bi omogočal ustrezen zajem. V letu 2004, ko število števcev z daljinskim odčitavanjem, zmožnih

odbiranja 15-minutnih vrednosti, še ni bilo tako veliko kot danes, je bila problematika nadomestnih obratovalnih diagramov obdelana v [6].

Odpiranje trga se je v letu 2007 nadaljevalo do te mere, da so vsi odjemalci (tudi gospodinjski) postali upravičeni do proste izbire dobavitelja električne energije. Cene energije so dobavitelji začeli oblikovati tržno, v obliki različnih paketov in jih niso več določale predpisane tarife. V letu 2011 so elektrodistribucijska podjetja iz matičnih podjetij tudi formalno izčlenila hčerinska podjetja za dobavo električne energije.

Podjetja za dobavo električne energije na podlagi napovedi večjih odjemalcev ter na podlagi sumarnih dnevnih diagramov za vse ostale odjemalce na področju posameznega distribucijskega podjetja – t.i. 'preostali diagram bilančne skupine' (te podatke jim je dolžan zagotoviti elektrodistributer), sestavijo napoved potrebne energije, ki jo bodo morali dobaviti v prihodnjem obdobju. Del te energije dobavitelji zagotovijo z nakupom skozi dolgoročnejše pogodbe, del pa skozi dnevno trgovanje na borzi električne energije. Koliko energije bodo dobavitelji zakupili vnaprej, koliko pa na dnevnem nivoju, je stvar njihove strategije, na katero pa zelo vplivajo cenovne razmere in trendi na trgu električne energije. Od tega, kako dobro znajo dobavitelji napovedati svoje potrebe po energiji in kako znajo optimirati nakup le te, je odvisna njihova poslovna uspešnost. Dnevni diagrami porabe energije imajo tako za dobavitelja velik pomen.

Vendar v tej nalogi ne bomo obravnavali dobaviteljevega vidika uporabe dnevnih obremenilnih diagramov, temveč se bomo osredotočili na distributerja in preprečevanje odtekanja prihodkov distributerja. Na podlagi oblike dnevnega obremenilnega diagrama bomo skušali razvrstiti odjemalce v tipične skupine in na tak način spoznati značilnosti tipičnih odjemalcev. Spremembe v obliki tipičnega dnevnega obremenilnega diagrama skozi čas lahko pomenijo tudi spremenjene navade oz. pojav novih vrst in načinov odjema v omrežju. Na podlagi tovrstnih informacij lahko distribucijsko podjetje bolje načrtuje razvoj svojega omrežja, da bo to prilagojeno bodočim potrebam. Odstopanja v obliki dnevnega diagrama posameznega odjemalca od tipičnega diagrama 'njegove' skupine lahko kažejo tudi na neupravičen odjem oz. krajo električne energije (odjem mimo števca). Tako imajo tipični dnevni diagrami velik pomen tudi za distributerja in to za uporabo v več vidikih.

3 Razvrščanje v skupine

Odjemalce bomo razvrstili v skupine tako, da bodo v skupine združeni tisti s podobno dnevno dinamiko porabe energije, oz. s podobnim normiranim dnevnim diagramom porabe energije.

Proces primerjanja in razvrščanja vzorcev v skupine (angl. clustering), ki je za človeški um nekaj povsem naravnega, saj to počne ves čas, je tudi v strojnem učenju osnovna disciplina. Kljub temu pa gre v splošnem za zelo zahteven problem. Raziskovalci so se ga lotevali na mnogo načinov, posledica česar je veliko število različnih postopkov in metod. Različne metode lahko dajo različne rezultate razvrstitve vzorcev. Katera razvrstitev je ustreznejša in posledično, katera metoda je primernejša, je odvisno tudi od množice vzorcev, katere razvrščamo, in od same problemske domene. Ob tem se hkrati pojavljajo vedno novi kriteriji ocenjevanja uspešnosti [12]. Cilj razvrščanja je poiskati skupine tako, da bodo vzorci v posamezni skupini čim bolj podobni, vzorci v različnih skupinah pa čim bolj različni. Pri razvrščanju odkrivamo medsebojne podobnosti, ugotavljamo povezave med vzorci, preiskujemo njihove strukture in raziskujemo znanje, skrito v podatkih. Zato področje razvrščanja v skupine spada pod okrilje širokega pojma razpoznavanja vzorcev (angl. pattern recognition) in podatkovnega rudarjenja (angl. data mining). Gre za nenadzorovan (angl. unsupervised) proces, kjer vnaprejšnje znanje o vzorcih ni poznano, kar pomeni, da ne poznamo pravil za razvrščanje, oz. skupine niso vnaprej določene. Nasprotno 'uvrščanje' (angl. classification) pomeni, da je število skupin in razvrstitev določenega števila vzorcev, ki predstavljajo učno množico, vnaprej poznano in zato predstavlja nadzorovano učenje.

3.1 Mere razdalj

Algoritmi za razvrščanje močno slonijo na merah podobnosti oz. razdaljah med vzorci, saj želimo razdaljo med elementi iste skupine minimizirati, razdaljo med elementi različnih skupin pa maksimizirati. Pri določitvi razdalje med dvema elementoma si pomagamo s pojmoma podobnost (angl. similarity) in neenakost (angl. dissimilarity). Obe meri sta preslikavi, ki vsakemu paru elementov priredita neko realno število.

$$\text{mera podobnosti} \quad s : (X, Y) \rightarrow R \tag{3.1}$$

$$\text{mera različnosti} \quad d : (X, Y) \rightarrow R \tag{3.2}$$

Mero podobnosti lahko pretvorimo v mero neenakosti (s katerokoli monotono padajočo transformacijo) in obratno (z monotono naraščajočo transformacijo).

Meri različnosti rečemo razdalja, kadar zadošča naslednjim pogojem:

$$\text{nenegativnost:} \quad d(X, Y) > 0 \tag{3.3}$$

$$d(X, X) = 0 \tag{3.4}$$

$$\text{simetričnost:} \quad d(X, Y) = d(Y, X) \tag{3.5}$$

$$\text{razločljivost:} \quad d(X,Y) = 0 \implies X = Y \tag{3.6}$$

$$\text{trikotniška neenakost:} \quad \forall Z: d(X,Y) \leq d(X,Z) + d(Z,Y) \tag{3.7}$$

X in Y sta D-dimenzionalna elementa množice, katero razvrščamo.

- razdalja Minkowskega: $dis(X,Y) = (\sum_{i=0}^{D} |x_i - y_i|^p)^{\frac{1}{p}}$ (3.8)

ali njeni posebni primeri:

- razdalja Manhattan ali »City-block« (p=1): $dis(X,Y) = \sum_{i=0}^{D} |x_i - y_i|$ (3.9)
- evklidska razdalja (p=2): $dis(X,Y) = \sqrt{\sum_{i=0}^{D} (x_i - y_i)^2}$ (3.10)
- razdalja Čebiševa (p=∞): $dis(X,Y) = \max_i |x_i - y_i|$ (3.11)

Obstaja še veliko drugih razdalj (npr.: kosinusna razdalja, Pearsonov korelacijski koeficient, Jeffreyjeva divergenca, razdalja Canberra, Hellingerjeva razdalja, razdalja Mahalanobisova, Jaccardov koeficient, …). Nekatere so primerne za vzorce z diskretnimi oz. tudi z binarnimi vrednostmi.

Pomembno je, da pred razvrščanjem danih vzorcev izberemo najustreznejšo mero, glede na naravo vzorcev in njihovo razporeditev.

3.2 Razvrščanje v skupine

Povsem jasno razvrstitev tehnik za razvrščanje v skupine je nekoliko težavno podati, saj te kombinirajo nekatere osnovne pristope oz. se prekrivajo v nekaterih osnovnih lastnostih.

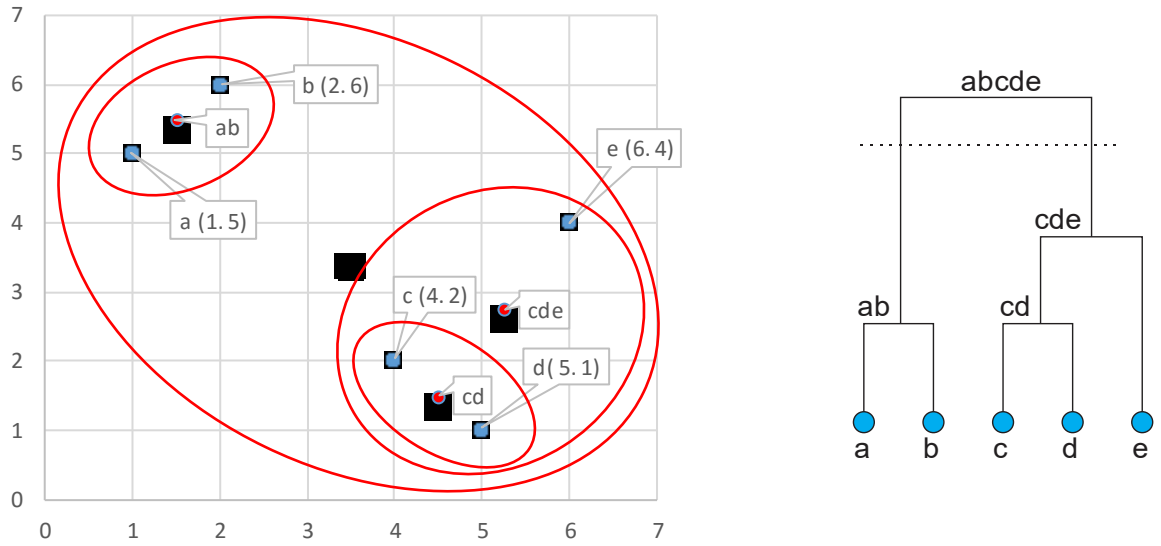
Vendarle pa jih v grobem lahko razdelimo na hierarhične (angl. hierarchical) in delitvene (angl. partitioning) [9, 18] ter na tehnike, temelječe na gostoti razporeditve (angl. density-based), mrežni razporeditvi (angl. grid-based) ter razporeditvi po modelu (angl. model-based).

3.3 Hierarhični postopki

S hierarhičnimi postopki zgradimo celo množico skupin, kjer je na najvišjem nivoju ena sama skupina, ki vključuje vse vzorce, na najnižjem nivoju pa je vsak vzorec sam v lastni skupini. To strukturo lahko gradimo od spodaj navzgor, z združevanjem (angl. agglomerative) ali od zgoraj navzdol, s cepitvijo (angl. divisive) [10 pog. 9.5]. Pri združevanju začnemo tako, da v prvem koraku vsakemu objektu dodelimo skupino. Vsaka skupina ima en element, razdalje med skupinami pa so enake razdaljam med objekti. V drugem koraku skupini, ki imata najmanjšo medsebojno razdaljo, združimo v novo skupino. V tretjem koraku izračunamo razdalje med novo skupino in vsako izmed starih skupin. Nato ponavljamo koraka 2 in 3, dokler vseh gruč ne združimo v eno samo. Različni algoritmi različno določajo razdaljo med skupinami. Ta je

lahko npr. razdalja med težišči skupin oz. med srednjimi vrednostmi vseh elementov v skupini, lahko je razdalja med najbližjima elementoma dveh skupin ali najbolj oddaljenima elementoma, lahko tudi med vsakim elementom iz prve in vsakim iz druge skupine. Rezultati združevanja so odvisni tako od izbrane preslikave za mero različnosti oz. za razdaljo, kot od algoritma določanja razdalje med skupinami. V vsakem primeru pa je vsak element množice v natanko eni skupini – noben ne more biti nerazvrščen, niti ne more biti v več skupinah hkrati.

Če ponazorimo hierarhično dekompozicijo združevanja v drevo tako, da je dolžina vej sorazmerna z razdaljo med skupinami, dobimo dendrogram. Če 'prerežemo' dendrogram na mestu, kjer so veje najdaljše, dobimo objekte, združene v naravno število skupin.



Slika 3.1: Primer množice s 5 2-razsežnimi elementi in dendrogram njihovega postopnega združevanja

Za primer na sliki 1, na podlagi hierarhičnega združevalnega razvrščanja množice petih elementov vidimo, da je optimalno število skupin 2 in da sta v eno skupino razvrščena elementa a in b, v drugo pa elementi c, d in e.

Hierarhične metode niso primerne za velike množice, saj s številom elementov N (in številom dimenzij) strmo narašča (v razmerju z N^3) potrebno število računskih operacij pri računanju razdalj.

3.4 Delitveni postopki

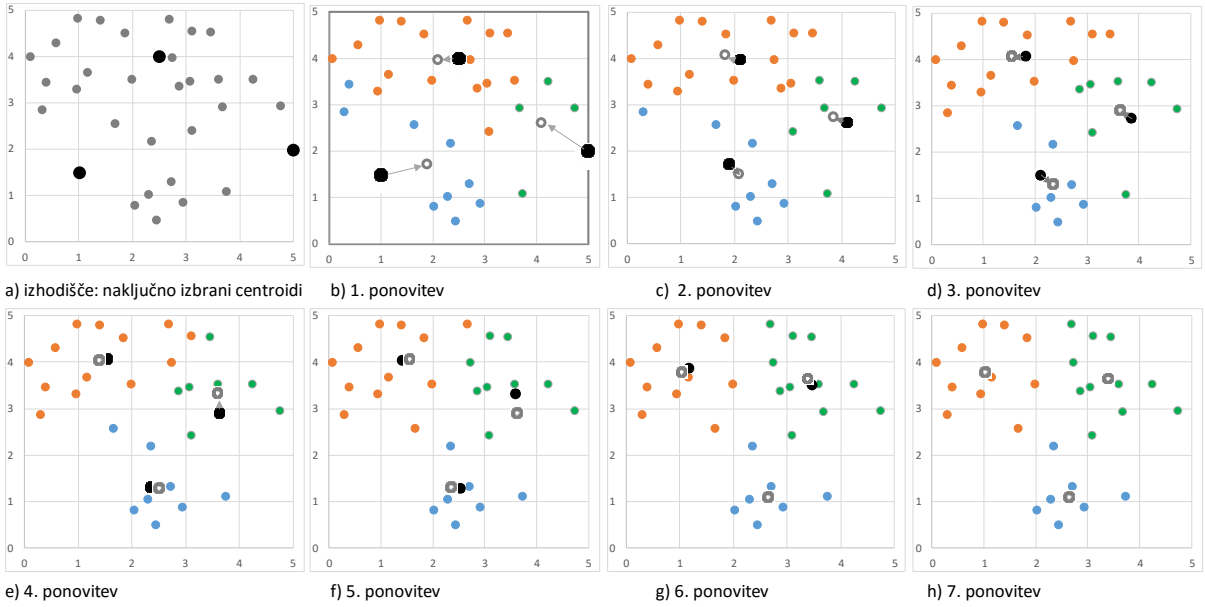
Delitveni postopki se ukvarjajo z izzivom, kako množico N-tih i-razsežnih vzorcev razporediti v K skupin, pri čemer optimizirajo čim večjo podobnost med elementi znotraj posamezne skupine in čim večjo različnost med elementi v različnih skupinah. Podlaga za različnost (in podobnost) je razdalja, katero določamo z različnimi preslikavami.

3.4.1 Metoda K-voditeljev (K-means)

Pogosto uporabljen algoritem za razvrščanje elementov v skupine je algoritem K-voditeljev (angl. K-means). Kot vhodni podatek moramo navesti K - število skupin, v katere želimo razvrstiti elemente množice. Algoritem najprej naključno določi K centroidov v prostoru, v katerem se nahajajo elementi naše množice in nato ponavlja naslednji postopek:

- vsak element v zbirki priredi najbližjemu centroidu (razdaljo določa izbrana preslikava, npr. evklidska razdalja), ter tako formira K skupin, v katere so razvrščeni vsi elementi,
- za vsako skupino izračuna center oz. povprečno vrednost elementov, ki jih vsebuje ter centroid prestavi na to novo središčno točko.

Postopek ponavlja, dokler se ob iteraciji še kakšen element premakne iz ene skupine v drugo, oz. dokler se spreminjajo pozicije centroidov.



Slika 3.2: Prikaz razvrščanja elementov po metodi K-voditeljev (za K=3)

Algoritem z ustrezno izbrano mero oddaljenosti vedno konvergira. Vendar je lahko, ob izbranih različnih začetnih lokacijah centroidov, različen tudi končni rezultat. Zato ni vseeno, kam postavimo začetne točke. Pri tem si lahko pomagamo z domenskim znanjem, s pomočjo katerega postavimo dodatne omejitve oz. pravila ali pa s predhodnim raziskovanjem razporeditve elementov množice.

Rezultat je seveda zelo odvisen tudi od števila skupin, katerega smo podali na začetku. Pri določitvi števila skupin, si zopet lahko pomagamo z domenskim znanjem s področja, za katerega razvrščamo vzorce ali pa postopek razvrščanja ponovimo za različno določena števila skupin in končne rezultate primerjamo tako, da npr. seštejemo razdalje vseh elementov do njihovih centroidov.

Posamezen element množice je pri algoritmu K-voditeljev vedno razvrščen v le eno skupino. Lahko pa se zgodi, da ostane kakšna skupina prazna. V takšnem primeru lahko centroid prazne skupine premaknemo npr. v bližino skupine, ki ima veliko vsoto razdalj in ponovimo razvrščanje.

Metoda s K-voditelji je priljubljena, saj je relativno enostavna in učinkovita. Ima pa nekaj slabosti:

- primerno število skupin K moramo poznati vnaprej,
- občutljiva je na začetno postavitev centroidov,
- občutljiva je na šum in osamelce,
- ni primerna za razporeditve elementov, ki niso krožne oblike ali so zelo različnih velikosti,
- razvrščamo lahko le elemente, ovrednotene z numeričnimi vrednostmi.

Z večanjem števila elementov množice, števila dimenzij in izbranega števila skupin, raste obseg potrebnega procesiranja približno linearno, kar je v primerjavi z nekaterimi drugimi metodami dokaj ugodno.

Razne variacije te metode zmanjšujejo ali odpravljajo nekatere njene slabosti. Tako npr. metoda s K-načini (angl. K-modes), ki uporabi novo mero različnosti, razširja uporabo algoritma tudi na diskretne vrednosti. Metoda s K-medoidi (angl. K-medoids) zmanjšuje težavo z osamelci. To stori tako, da centrov skupin ne premika na srednjo vrednost pripadajočih elementov, temveč center postavi na tistega od dejanskih elementov, ki je najbližji srednji vrednosti članov.

3.5 Tehnike temelječe na modelih (verjetnostni pristop k razvrščanju)

Tehnike, temelječe na modelih skušajo optimizirati ujemanje podanih vzorcev z določenim matematičnim modelom. Te metode se pogosto opirajo na domnevo, da so podatki razporejeni po neki kombinaciji verjetnostnih porazdelitev, običajno po mešanici Gaussovih verjetnostnih porazdelitev [4], [10, pog. 8.4, 9.6]. Problem, ki ga želimo rešiti je, kako nastaviti parametre verjetnostnih razporeditev, da se bodo te najbolje prilegale podatkom.

3.5.1 Metoda maksimiranja pričakovanj (EM - Expectation-Maximization)

Algoritem maksimiranja pričakovanj EM (Expectation-Maximization) lahko obravnavamo kot nadgradnjo algoritma K-voditeljev. Podobno kot pri K-voditeljih posamezen element uvrstimo v tisto skupino, centroidu katere je najbližji (glede na izbrano mero razdalje), pri EM elementu pripišemo verjetnosti, da pripada posamezni skupini. Posamezen primerek tako hkrati pripada več skupinam, vsaki z različno verjetnostjo, katerih vsota je 1. Za računanje verjetnosti EM uporabi Gaussovo normalno porazdelitev, katero za D-razsežni vzorec x zapišemo tako:

$$N(x|\mu,\sum) = \frac{1}{\sqrt{(2\pi)^D\,|\sum|}}\, exp\left\{-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\right\} \tag{3.12}$$

Kjer je μ D-razsežni vektor srednjih vrednosti, ∑ kovariančna matrika velikosti D × D, |∑| njena determinanta in oznaka ^{T} transpozicija vektorja. Od izbire kompleksnosti ∑ je odvisna tudi kompleksnost nastavljanja iskanih parametrov. Hkrati pa s kompleksnejšo obliko lahko bolje zajamemo določeno strukturo v podatkih, ki je sicer ne bi. Zadostno kompleksnost dosežemo z linearno kombinacijo (mešanico) enostavnih Gaussovih porazdelitev (angl. mixture of Gaussians). Posamezna enostavna Gaussova porazdelitev oz. komponenta mešanice v tej linearni kombinaciji predstavlja skupino, v katero želimo razporediti elemente. Vsaka komponenta ima svoj vektor μ in matriko Σ.

Superpozicijo K Gaussovih porazdelitev zapišemo:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{3.13}$$

Koeficiente mešanja (angl. mixture coefficients) označimo z π_k in velja:

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{in} \quad 0 \leq \pi_k \leq 1 \tag{3.14}$$

Koeficiente mešanja lahko razumemo kot verjetnost p(k), da bo izbrana k-ta komponenta mešanice, $N(x|\mu_k, \Sigma_k)$ pa kot pogojno verjetnost $p(x|k)$ vzorca x, če smo izbrali k-to komponento mešanja. Uvedemo še diskretne spremenljivke s_i, ki določajo pripadnost posameznega vzorca x določeni skupini. Torej $s_i=k$ pomeni, da x_i pripada skupini k. Potem drži, da

$$\pi_k = p(s = k) \tag{3.15}$$

Spremenljivke s_i imenujemo skrite spremenljivke (angl. latent, hidden variables). Naš cilj je, da v postopku optimizacije kriterijske funkcije, najdemo njihovo vrednost.

Oblika mešanice Gaussovih porazdelitev je torej odvisna od treh množic parametrov: $\pi = \{\pi_1, \dots, \pi_K\}$, $\mu = \{\mu_1, \dots, \mu_K\}$, $\Sigma = \{\Sigma_1, \dots, \Sigma_K\}$. Ena možnost za njihovo nastavitev je uporaba funkcije največje verjetnosti (angl. maximum likelihood). Sestavljena funkcija največjih verjetnosti je funkcija vzorcev ($x \in P$) in parametrov modela (Θ) in je definirana s:

$$p(P|\Theta) = p(\{x_1, x_2, \dots, x_N\}|\pi, \mu, \Sigma) = \prod_{i=1}^{N} p(x_i|\, \pi, \mu, \Sigma) \tag{3.16}$$

Cilj je poiskati parametre $\Theta = \{\pi, \mu, \Sigma\}$ tako, da ti maksimirajo logaritem verjetnostne funkcije (angl. log likelihood function). Za množico vzorcev $P = \{x_1, x_2, \dots x_N, \}$ in z upoštevanjem (3.16) je logaritem funkcije L enak

$$L = Ln\, p(P|\pi, \mu, \Sigma) = \sum_{i=1}^{N} \ln \{\sum_{k=1}^{K} \pi_k\, N(x_i|\pi_k, \mu_k, \Sigma_k)\} \tag{3.17}$$

Analitična rešitev verjetnostne funkcije zaradi logaritma notranje vsote ni možna, zato iščemo rešitev na drugačen način. Za iskanje verjetnostne funkcije s skritimi spremenljivkami bomo uporabili algoritem maksimiranja pričakovanja ali EM.

Najprej je potrebno L odvajati po srednji vrednosti μ_k in odvod postaviti na 0. Dobimo:

$$\mu_k = \frac{\sum_{i=1}^{N} r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}} \tag{3.18}$$

Kjer je r_{ik} posteriorna verjetnost pripadnosti vzorca x_i k-ti komponenti mešanice Gaussovih funkcij, kar lahko zapišemo kot $p(s_i = k|x_i, \mu_k, \Sigma_k)$ in se po Bayesovem teoremu izračuna kot:

$$r_{ik} = \frac{\pi_k N(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j,\Sigma_j)} \tag{3.19}$$

Podobno kot prej, odvajamo L po $\sum_k^{-1}$ in odvod postavimo na 0. Dobimo enačbo za $\sum_k$, ki se glasi:

$$\Sigma_k = \frac{\sum_{i=1}^{N} r_{ik}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{N} r_{ik}} \tag{3.20}$$

Sedaj moramo le še odvesti logaritem verjetnostne funkcije L po π_k. Pri tem moramo upoštevati omejitev, da je $\sum_{k=1}^{K} \pi_k = 1$. To dosežemo z uporabo Lagrangovega multiplikatorja in maksimizacije izraza $L' = L + \lambda(1 - \sum_{k=1}^{K} \pi_k)$. L'torej odvajamo po π_k in odvod izenačimo z 0. Po izračunu dobimo:

$$\pi_k = \frac{\sum_{i=1}^{N} r_{ik}}{N} \tag{3.21}$$

Število $N\pi_k$ lahko interpretiramo kot efektivno število vzorcev v skupini C_k.

Sedaj imamo vse pripravljeno za razlago delovanja algoritma EM. Bistvena sta dva koraka, E (pričakovanje) in M (maksimiranje), po katerih je algoritem tudi dobil ime. V koraku E izračunamo vrednost skritih spremenljivk, če so parametri modela Θ nespremenljivi, fiksni. V koraku M pa popravimo parametre glede na vrednost skritih spremenljivk. V primeru modela mešanice Gaussovih porazdelitev so s_i skrite spremenljivke, parametri Θ pa π_k, μ_k in Σ_k.

Algoritem EM za model mešanice Gaussovih porazdelitev po korakih:

Korak 1: Nastavimo začetne vrednosti parametrom π_k, μ_k in Σ_k za vsako komponento k=1, 2, …, K. Izračunamo začetno vrednost funkcije L (po enačbi 3.17).

Korak 2 (Korak E): Izračunamo vrednost r_{ik} (po enačbi 3.19). Uporabimo trenutno nastavljene parametre π_k, μ_k in Σ_k.

Korak 3 (Korak M): Izračunamo nove vrednosti parametrov μ_k^{novi}(po enačbi 3.18), Σ_k^{novi}(po enačbi 3.20) in π_k^{novi}(po enačbi 3.21). Uporabimo vrednosti r_{ik} izračunane v koraku 2.

Korak 4: Izračunamo vrednost logaritma verjetnostne funkcije L (po enačbi 3.17) in preverimo konvergenco bodisi parametrov, bodisi funkcije L. V primeru, da konvergenca ni dosežena, ponovimo postopek od koraka 2 dalje, sicer končamo.

V koraku 1 je mogoče nastaviti parametre μ_k z uporabo nekaj iteracij algoritma K-voditeljev, kar da boljše izhodišče, kot naključna izbira. Začetni π_k nastavimo na $\frac{1}{K}$, Σ_k pa je navadno enotska matrika ali njen ekvivalent.

Pri izvajanju algoritma lahko pride do situacije, da dobi j-ta komponenta Gaussove mešanice srednjo vrednost μ_j enako, kot je pozicija vzorca x_n, torej velja $\mu_j = x_n$. V limiti, ko gre $\sigma_j \rightarrow 0$, gre funkcija L proti neskončnosti. Zato maksimizacija funkcije L ni več rešljiv problem, kar

imenujemo problem singularnosti. Da se takim primerom izognemo, je potrebno uporabiti hevristične postopke. Ena izmed možnosti je, da ob zaznani singularnosti srednjo vrednost μ kritične komponente mešanja ponastavimo na neko naključno vrednost, kovariančno matriko iste komponente pa na neko veliko vrednost. Nato nadaljujemo z optimizacijo.

3.6 Klasifikacija - uvrščanje v predpisane skupine

Cilj pri klasifikacijskih postopkih je iz učne množice (tj. elementov, ki so že razvrščeni v znane skupine), odkriti lastnosti, ki so v korelaciji z razvrstitvijo. S tem pridobimo razumevanje o obnašanju množice. Tega uporabimo v drugi fazi, pri razvrščanju novih primerov v vnaprej predpisane skupine. Rezultat prve faze, v kateri iz učne množice pridobimo pravila za uvrščanje, je lahko predstavljen v obliki odločitvenih dreves (angl. decision trees) ali z množico klasifikacijskih ali povezovalnih pravil (angl. classification rules, association rules)

3.6.1 Odločitvena drevesa

Odločitvena drevesa predstavljajo enostaven način analize več spremenljivk, s katerim lahko opišemo znane elemente, kot tudi vrednotimo ali klasificiramo nove. Uporabljamo jih lahko na različne načine. Klasifikacija pri podatkovnem rudarjenju je eden od njih. Kadar je odvisna spremenljivka diskretna, govorimo o klasifikacijskih oz. razvrščevalnih drevesih, kadar je odvisna spremenljivka zvezna, pa o regresijskih drevesih. Drevo predstavimo s hierarhično drevesno strukturo, sestavljeno iz vozlišč, ki predstavljajo teste atributov (tj. vprašanja za odločanje) in usmerjenih povezav, ki predstavljajo odgovore oz. odločitve. Končna vozlišča drevesa (listi) predstavljajo končne razrede, v katere so elementi na koncu razvrščeni. Testiranje primerkov se vedno prične v korenu drevesa in konča v listih. Pot, po kateri gre testirani vzorec, in na katerem listu jo bo končal, je odvisna od vrednosti atributov, ki jih testiramo v vmesnih vozliščih.

Odločitveno drevo izdelamo z indukcijskim algoritmom. Poznamo različne indukcijske algoritme. Ti gradijo drevo od debla proti listom tako, da se v vsakem koraku vprašajo, kateri je najbolj informativni atribut in nato tega uporabijo za nadaljnjo klasifikacijo.

Algoritme, ki gradijo drevo z iskanjem kombinacij s križanjem in mutacijami atributov ter iterativne izgradnje drevesa z izboljševanjem, imenujemo evolucijski.

3.6.2 Povezovalna pravila

Povezovalna pravila (angl. association rules) omogočajo odkrivanje razmerij oz. povezav med atributi v opazovani množici primerov. Prav tako kot odločitvena drevesa, sodijo v opisno odkrivanje zakonitosti, saj rezultate uvrščanja lahko pojasnimo. Predstavljena so z:

$$X \rightarrow Y \text{ (zanesljivost, podpora)}$$

Zanesljivost predstavlja verjetnost, da se ob dogodku X zgodi tudi dogodek Y. Podpora je verjetnost, da X in Y nastopita hkrati.

Povezovalna pravila so pogosto uporabljena v analizah nakupovalne košarice, ki ugotavljajo, kateri izdelki se prodajajo skupaj.

Klasifikacijska pravila (angl. classification rules) so podobna povezovalnim pravilom, s to razliko, da v sklepnem delu nastopa samo ena postavka, ki je razred (saj so namenjena klasifikaciji).

4 Zasnova platforme za podporo preprečevanju odtekanja prihodkov v elektrodistribucijskem podjetju

V praktičnem delu naloge smo zasnovali in izdelali sistem, ki bo podprl proces preprečevanja odtekanja prihodkov v elektrodistribucijskem podjetju. Izveden je v testnem okolju Elektra Celje, d. d., podjetja za distribucijo električne energije. Uporabljeni so dejanski podatki, vendar skladno z dovoljenjem podjetja le tisti del podatkov, ki ne razkriva identitete strank. Osnovna ideja je bila izdelati analitsko – poročevalsko rešitev, ki bo združevala podatke iz različnih informacijskih sistemov, ki so vpleteni v procesih, vezanih na obračun omrežnine. Rešitev naj omogoči izvajanje sprotnih poizvedb naprednejšim uporabnikom – analitikom, hkrati pa naj omogoči poročanje v obliki predpripravljenih standardnih poročil. Glavni cilj je pripraviti analitsko strukturo, s katero bo možno iz velike množice podatkov števčnih meritev izluščiti informacijo o dinamiki odjema in na ta način odkriti tipične skupine odjemalcev ter vanje razvrstiti vse odjemalce. Še posebej nas zanimajo tisti, ki odstopajo od tipičnih vzorcev.

Eden od izzivov je, kako obdelati veliko količino podatkov, ki izvirajo iz sistema naprednega merjenja (AMI) in iz njih izluščiti koristne informacije. V tabelah, ki hranijo podatke o števčnih meritvah in stanjih, je bilo v času, ko smo pristopili k izdelavi, več kot 10^{10} zapisov.

4.1 Opis infrastrukture

Kot platformo za izvedbo naloge smo izbrali programski paket Microsoft SQL Server 2014 Enterprise, kakršen je tudi sicer v uporabi v Elektru Celje – kot mehanizem relacijske podatkovne baze in kot skupek storitev, ki omogočajo poslovno inteligenco (BI – Business Intelligence). Ta paket vsebuje več komponent [2], od katerih smo uporabili naslednje:

Sistem upravljanja podatkovnih baz SQL strežnik (SQL Server Engine); je storitev, ki omogoča hranjenje, obdelavo in varovanje relacijskih podatkovnih baz.

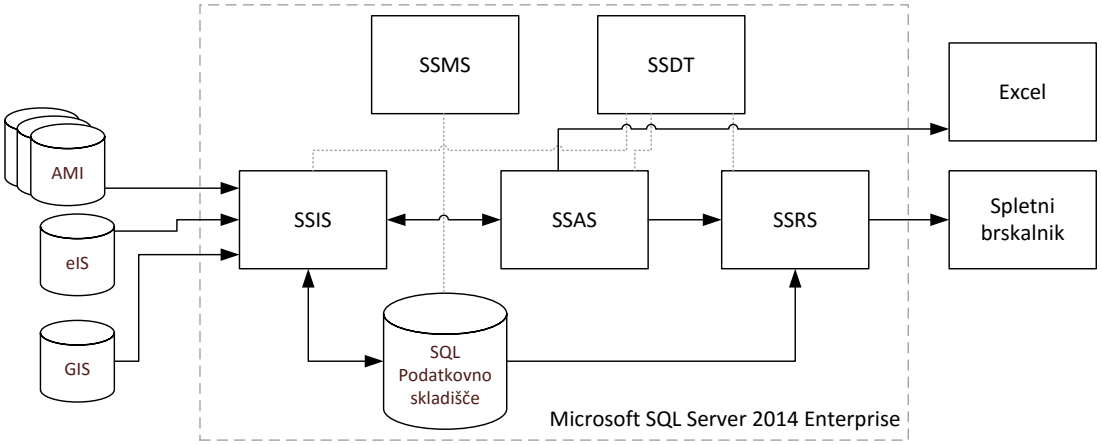
Analitične storitve strežnika SQL (SSAS - SQL Server Analysis Services); je storitev, ki omogoča vzpostavitev večdimenzionalne analitične strukture (OLAP – Online Analytical Processing) in izvajanje različnih algoritmov podatkovnega rudarjenja.

Integracijske storitve strežnika SQL (SSIS - SQL Server Integration Services); je storitev, s katero izvajamo zajemanje, preoblikovanje in polnjenje podatkov (ETL – Extract-Transform-Load) ter avtomatiziramo procesiranje analitskih struktur in modelov.

Poročevalske storitve strežnika SQL (SSRS – SQL Server Reporting Services); so storitve, s katerimi predpripravljena dinamična poročila spletno ponudimo uporabnikom.

Podatkovna (razvojna) orodja strežnika SQL (SSDT – SQL Server Data Tools); je razvojno orodje, ki temelji na ogrodju Microsoft Visual Studio, a je prilagojeno za razvoj analitskih struktur, modelov podatkovnega rudarjenja ter za izdelavo procesov ETL.

Konzola za upravljanje strežnika SQL (SSMS - SQL Server Management Studio); je grafični uporabniški vmesnik za upravljanje mehanizma podatkovnih baz, analitskih storitev, repozitorija integracijskih storitev, mehanizma poročevalskih storitev ter za izvajanje poizvedb SQL, MDX (angl. Multidimensional Expressions), DMX (angl. Data Mining Extensions) in XMLA (angl. XML for Analysis).



Legenda:

AMI – Sistem naprednega merjenja (Advanced Metering Infrastructure)
GIS – Prostorski informacijski sistem (Geographic Information System)
eIS - obračunski sistem elektrodistribucij Slovenije

SQL - Sistem upravljanja podatkovnih baz (SQL Server Engine)
SSAS - Analitične storitve SQL strežnika (SQL ServerAnalysis Services)
SSIS - Integracijske storitve SQL strežnika (SQL Server Integration Services)
SSRS - Poročevalske storitve SQL strežnika (SQL Server Reporting Services)
SSMS - Konzola za upravljanje SQL strežnika (SQL Server Management Studio)
SSDT - Podatkovna (razvojna) orodja SQL strežnika (SQL Server Data Tools)

Slika 4.1: Gradniki poskusnega sistema

Vse naštete komponente strežnika SQL so nameščene na enem strežniku z operacijskim sistemom Microsoft Windows Server 2012 R2 Datacenter, ki je nameščen na navideznem strežniku z dodeljenimi naslednjimi viri: 4 logični procesorji (2,7 GHz), spomin RAM 128 GB, prostor na diskovnem polju: 2TB (na eni logični enoti). Diskovno polje je srednjega cenovnega razreda, IBM storwize v7000. Virtualizacijsko okolje, v katerem teče strežnik, je VMware vSphere 6.0.

Celotno okolje smo tako zasnovali na povsem klasični platformi za poslovno inteligenco, brez uporabe ogrodij z razpršenim procesiranjem, namenjenih obdelavi zelo velikih količin podatkov (ti. 'Big Data'). Med drugim smo želeli preizkusiti ali naš primer z obstoječo količino podatkov v takšnem okolju naleti na kakšne resne omejitve zmogljivosti sistema. Za postavitev v produkcijskem okolju bi analitsko (SSAS) in poročevalsko (SSRS) storitev zagotovo ločili na samostojna strežnika. Hkrati bi tudi razmišljali o arhitekturi visoke razpoložljivosti. Pa tudi za povečanje zmogljivosti z dodajanjem dodatnih virov in s skalabilnostjo sistema je še veliko možnosti.

4.2 Priprava podatkovnega skladišča

Pri modeliranju podatkovnega skladišča smo se v prvi iteraciji omejili na mere in dimenzije za katere smo menili, da so najpomembnejše pri zasledovanju razlik med izmerjenimi in obračunanimi količinami ter pri raziskovanju odjemalcev glede na dinamiko porabe električne energije. V naslednjih iteracijah bi bilo smiselno skladišče vsebinsko razširiti.

Zasnova podatkovnega skladišča je povsem klasična zvezdna struktura, z nekaj tabelami dejstev in denormaliziranimi dimenzijskimi tabelami, ki ta dejstva opisujejo. Podatkovno bazo skladišča smo poimenovali s kratico EDW (kot: elektrodistribucijsko podatkovno skladišče). Njegov model prikazuje slika 4.2.

Glavna vira podatkov sta dva in sicer:

eIS: informacijski sistem za obračun električne energije in omrežnine ter spremljanje 'življenjskega cikla' odjemalca. Sistem je po naročilu slovenskih elektrodistribucijskih podjetij njim in pravilom slovenskega trga z električno energijo 'na kožo' razvilo podjetje Informatika d. d.. Ta sistem zagotavlja podporo za vse procese, naštete v točki 2.2.2. – torej poleg samega obračuna vključuje tudi izdajanje elektroenergetskih soglasij, pogodb o priključitvi, pogodb o dostopu, vodenje evidence merilnih (odjemnih) mest, števcev ter tudi spremljanje plačil z modulom masovnih saldakontov ter izvajanja procesa izterjave.

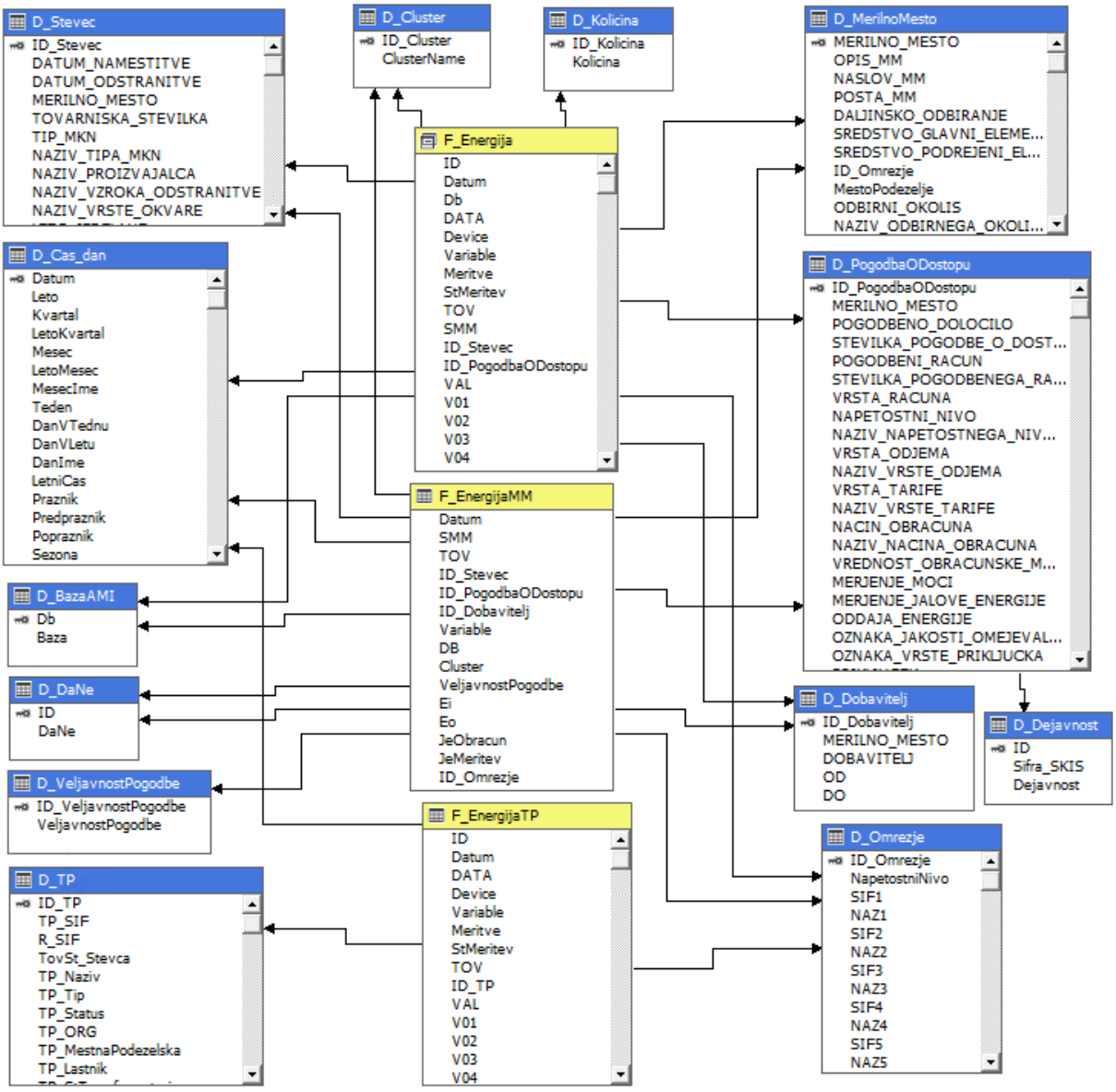
AMI (Advanced Metering Infrastructure): sistem naprednega merjenja 'Advance' proizvajalca Enerdat-S, ki se trži pod znamko podjetja Landis & Gyr. V Elektru Celje je sistem zastavljen tako, da podatke, zbrane s pametnih števcev, zapisuje v tri podatkovne baze – ločeno za večje industrijske odjemalce, za gospodinjske in manjše poslovne odjemalce ter za meritve v energetskih objektih (v transformatorskih postajah).

Tretji vir, ki smo ga uporabili, je prostorski informacijski sistem **GIS** (Geographic Information System) ArcGIS, proizvajalca ESRI, nadgrajen z branžno rešitvijo Arc FM UT (angl. Facility Management for Utilities), proizvajalca Sicad. Iz tega sistema pridobimo podatke o povezanosti elementov v omrežju ter o namestitvi števcev v transformatorskih postajah.

Večino dimenzij smo načrtovali tako, da so časovno spremenljive (tipa 2), kar pomeni, da če se je nek pojavek v zgodovini spremenil, imamo v dimenzijski tabeli zapisano eno vrstico s stanjem pred spremembo in vpisanim datumom spremembe kot mejnik, do katerega ta podatek velja ter drugo vrstico s stanjem po spremembi.

Pri vseh virih smo se srečali z izzivi, kako prečistiti podatke. Na več mestih smo namreč naleteli na nekonsistentnosti. Nekatere od teh so takšne narave, da vzrokov zanje ne moremo povsem odpraviti (npr. neveljavni odčitki zaradi okvare števcev) in bodo torej nastajale tudi v prihodnje. Za takšne moramo najti pravilo, da jih, kadar jih zaznamo, povsem popravimo (če je to mogoče, npr. aproksimiramo manjkajočo vrednost) ali pa jih zaradi nepravilnosti izločimo iz nadaljnje obravnave. Drugi primeri so takšne narave, da nastajajo zaradi pomanjkljivih omejitev oz. preohlapnih pravil v uporabniških aplikacijah, kjer ti podatki nastajajo ali kar je najpogosteje, pri integracijah med sistemi. V takšnih primerih moramo identificirati pravi vzrok, ki dopušča

nastajanje nepravilnosti in ga najprej odpraviti na viru (npr. popraviti aplikacijo ali spremeniti pravila za izvajanje poslovnega procesa) in nato še popraviti podatke. Tretji primer so nepravilnosti, ki so nastale zaradi kakšne od migracij v preteklosti in ne nastajajo več. V teh primerih lahko le popravimo izvorne podatke, če je to možno, če ne pa skušamo anomalijo ustrezno upoštevati v ETL postopku za polnjenje podatkovnega skladišča. Nepravilnosti, na katere smo naleteli ob izgradnji podatkovnega skladišča, bomo obravnavali v nadaljevanju, pri vsaki dimenziji posebej.



Slika 4.2: Model podatkovnega skladišča EDW

4.2.1 Tabele dejstev

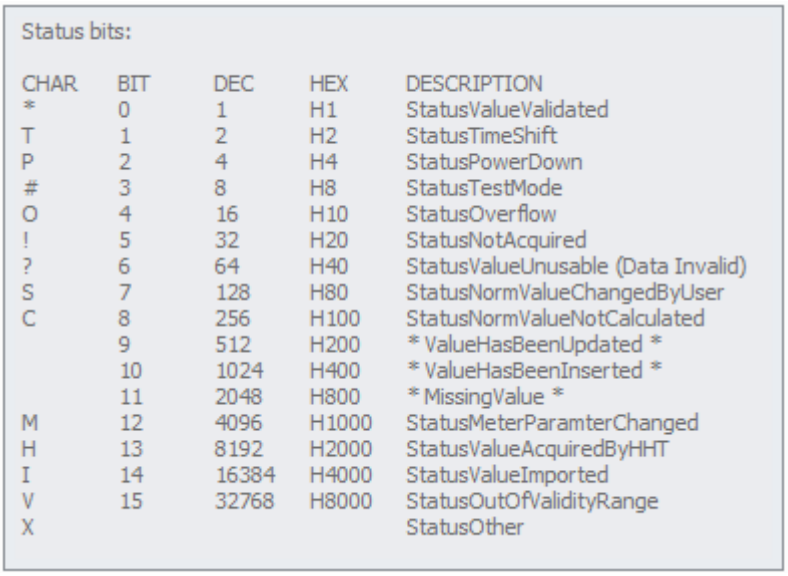
V tabele dejstev, ki so v modelu, na sliki 4.2 označene z rumeno barvo, hranimo mere, katere bomo analizirali in medsebojno primerjali ter tuje ključe, preko katerih se povežemo z dimenzijskimi tabelami. Kot je omenjeno v prejšnji točki, smo se po razmisleku odločili, da se v prvi iteraciji izgradnje analitične strukture, ki je predmet te naloge, osredotočimo le na količine energije in ne na druge storitve, ki jih podjetje izvaja in tudi ne na zneske in finančno plat. Za celovito preprečevanje odtekanja prihodkov bi bilo smiselno v naslednjih iteracijah izgradnje dodati tudi te vsebine. Vsekakor pa bi sedaj izbrane vsebine morale zadoščati za tisti del, kjer pričakujemo največje učinke in za potrditev koncepta za analizo obsežnih števčnih podatkov. V podatkovnem skladišču smo tako zasnovali tri tabele dejstev, z naslednjimi vsebinami:

4.2.1.1 Dnevni diagrami odjemalcev

V tabeli 'F_Energija' so podatki o količini energije, ki jo odjemalci prejmejo iz omrežja na merilnih mestih, opremljenih s pametnimi števci, vključenimi v sistem naprednega merjenja. Gre za podatke o pretokih (oddaji (+) in prejemu (-)) delovne (A) in jalove (R) energije (dobimo vrste: A+, A-, R+ in R-). Kot smo nakazali že v točki 2.3, pametni števec zajema veliko različnih veličin in dogodkov ter jih hrani v svojih registrih. Vsebino nekaterih izmed registrov redno (običajno enkrat dnevno) s sistemom naprednega merjenja (AMI) prenesemo in prepišemo v osrednjo podatkovno bazo. Med te registre sodijo prej omenjene vrste energij, zajete za 15-minutne intervale. To pomeni 4x24=96 zapisov na dan za eno merilno mesto, za vsako vrsto energije. Ob tem vsaj enkrat dnevno zajamemo tudi števčna stanja registrov za posamezne tarife (visoke – VT, manjše – MT in enotne – ET) ter zapise o posebnih dogodkih. Za segment večjih industrijskih odjemalcev s 15-minutnimi intervali zajemamo tudi tokove in napetosti po fazah ter različne dosežene dnevne konice moči.

Vsak odčitek v sistemu AMI, ki govori o vrednosti določenega registra za določen števec, je opremljen s časovno značko in s statusom. Status je predstavljen s celoštevilčnim številom, pri čemer ima vsak bit tega števila (v dvojiškem zapisu) svoj pomen, kot ga prikazuje slika 4.3.

Pri prenosu v podatkovno skladišče upoštevamo le zapise, ki nimajo postavljenih bitov 5 (StatusNotAcquired) ali 6 (StatusValueUnusable-Data Invalid).



Status bits:

CHAR	BIT	DEC	HEX	DESCRIPTION
*	0	1	H1	StatusValueValidated
T	1	2	H2	StatusTimeShift
P	2	4	H4	StatusPowerDown
#	3	8	H8	StatusTestMode
O	4	16	H10	StatusOverflow
!	5	32	H20	StatusNotAcquired
?	6	64	H40	StatusValueUnusable (Data Invalid)
S	7	128	H80	StatusNormValueChangedByUser
C	8	256	H100	StatusNormValueNotCalculated
	9	512	H200	* ValueHasBeenUpdated *
	10	1024	H400	* ValueHasBeenInserted *
	11	2048	H800	* MissingValue *
M	12	4096	H1000	StatusMeterParamterChanged
H	13	8192	H2000	StatusValueAcquiredByHHT
I	14	16384	H4000	StatusValueImported
V	15	32768	H8000	StatusOutOfValidityRange
X				StatusOther

Slika 4.3: Pomen posameznih bitov v atributu 'status' pri meritvah pametnega števca

Časovno značko pametni števci in sistem AMI vodijo po naravni (zimski) uri (z UTC+1) in ne sledijo premikanju ure na poletni čas. Zaradi tega moramo ob prenosu v podatkovno skladišče upoštevati premik ure v poletnem času.

Ob razmisleku, kako zapisati te podatke v podatkovno skladišče, smo se odločili, da jih bomo pretvorili tako, da bo en zapis predstavljal dnevni pretok energije, ločeno za vsako smer posebej ter za delovno in jalovo energijo. Količino posamezne vrste energije za dan izračunamo tako, da seštejemo vse odčitane vrednosti (96 vrednosti) tega dne. Bistvena pa je tudi informacija o dnevni dinamiki porabe v obliki 15-minutnih vrednosti. Po tej dinamiki želimo primerjati merilna mesta med seboj, ustvariti skupine s tipično obliko in merilna mesta razvrstiti vanje. Za uspešno primerjanje in razvrščanje pa moramo višino dnevnih krivulj poenotiti. To naredimo tako, da vrednost posameznega odčitka normiramo na dnevno vrednost. Ker želimo posamezno normirano vrednost zapisati s predznačeno 2-zlogovno celoštevilčno vrednostjo (oblika smallint), dobljene vrednosti pomnožimo z 10^4 in zanemarimo neceli del. Na tak način dobimo 96 normiranih vrednosti, ki jih kot ločene atribute dodamo v zapis o dnevni porabi. Ta del zapisa si lahko predstavljamo kot vektor

$$\mathrm{V}\ (96);\ \mathrm{v}_i = 10^4 \frac{x_i}{\sum_{k=1}^{96} x_k} \tag{4.1}$$

ki bo predstavljal element množice v 96-razsežnem prostoru, katero bomo razvrščali v skupine.

V eni vrstici naše tabele dejstev F_Energija je tako ob podatku o datumu in tem, za katero merilno mesto in za katero vrsto energije gre, zapisan še atribut z dnevno porabo energije (realno število) ter 96 celoštevilčnih atributov z vrednostmi med 0 in 10^4. Takšna oblika zapisa nam bo omogočala hitro obdelavo v fazah podatkovnega rudarjenja in analize.

Med števci v sistemu naprednega merjenja je žal tudi znaten delež starejših tipov, ki niso zmožni zajema 15-minutnih, ampak le urnih vrednosti. Zapisi za merilna mesta s takšnimi

števci imajo v atributih naše tabele, ki predstavljajo normiran dnevni diagram, izpolnjene le vrednosti v poljih, ki predstavljajo polne ure, medtem ko so vrednosti ostalih atributov 0.

Da bomo znali enostavno ločiti med temi vrstami zapisov, smo dodali še en atribut, v katerega zapišemo število meritev za izbrano merilno mesto, vrsto energije in dan. Vrednost tega atributa je torej tipično 96, za merilna mesta s števci z omejeno funkcionalnostjo pa 24. V primerih, ko izpadejo posamezne meritve, pa je ta vrednost tudi manjša od tipične. Ta atribut nam daje možnost enostavnega izločanja nepopolnih vzorcev.

Za eno merilno mesto imamo tipično 4 zapise za vsak dan (odtekajočo delovno energijo A+, pritekajočo delovno energijo A-, odtekajočo jalovo energijo R+, pritekajočo jalovo energijo R-).

4.2.1.2 Dnevni diagrami pretokov v transformatorskih postajah

Elektro Celje je v sklopu projekta, delno financiranega z evropskimi sredstvi, pametne števce vgradil tudi v transformatorske postaje, na nizkonapetostne izvode. Tako je sedaj s števci, ki so vključeni v sistem naprednega merjenja, opremljenih dve tretjini vseh nizkonapetostnih izvodov iz transformatorskih postaj. Podatek o količinah energije, ki je bila dostavljena v manjši segment omrežja, kot je posamezen izvod transformatorske postaje, je pri odkrivanju izgub (tehničnih in komercialnih) zelo dobrodošel, saj lahko načeloma neposredno primerjamo količino energije, ki je bila dostavljena v ta segment preko transformatorske postaje in tisto, ki je bila v tem segmentu proizvedena, z vsoto tiste, ki je bila dostavljena (in izmerjena ter kasneje obračunana) odjemalcem. Zato je povsem naravno, da smo se odločili tudi te podatke dodati v podatkovno skladišče. Zapisani so v tabeli F_EnergijaTP. Struktura zapisa je praktično enaka kot v tabeli F_Energija, opisani v prejšnji točki (4.2.1.1). Prav tako kot pri F_Energija imamo tudi v F_EnergijaTP mero, ki je vsota 15-minutnih količin energije v dnevu, torej predstavlja dnevno količino energije. Ob tem imamo 96 atributov, v katerih so normirane vrednosti, v obsegu vrednosti od 0 do 10^4. En zapis predstavlja energijo za en števec za en dan.

Razlika v primerjavi s pametnimi števci, ki so nameščeni pri odjemalcih je, da števci v transformatorskih postajah beležijo več parametrov, katere tudi prenašamo in zapisujemo v bazo sistema AMI. Ti dodatni parametri so predvsem takšni, da podajajo kakovost dobavljene energije in nam pomagajo pri tehnični optimizaciji omrežja. Za prenos tovrstnih informacij v podatkovno skladišče, se za sedaj nismo odločili.

4.2.1.3 Obračunane količine energije

Nasproti podatkom o izmerjenih količinah električne energije želimo primerjati količine, ki so bile skozi mesečne račune obračunane odjemalcem. S tem namenom smo v skladišču dodali tabelo dejstev F_EnergijaMM. Vir teh vsebin je obračunski sistem eIS. Pripravljen vpogled (view) na nivoju podatkovne baze, namenjen analizi, nudi podatke o posameznih obračunanih količinah po merilnih mestih. Vsak zapis ima določeno obdobje (z dnem začetka in konca obdobja), na katerega se nanaša obračun. Večinoma so to obdobja enega meseca, pri akontacijskih računih pa se vsaj enkrat letno pojavi še poračun, ki je narejen za daljše obdobje

in se prekriva z obdobji akontacijskih računov. Prekrivanja obračunskih obdobij za posamezno merilno mesto se dogajajo tudi iz drugih razlogov, kot npr. poračun zaradi spremembe pogodbe, spremembe plačnika, ugotovljen neupravičen odjem ali napaka v merjenju in podobno. Kadar gre za stornacijo nekega obračuna, je ta vedno kompletna. To pomeni, da postavka stornacije vedno v celoti kompenzira postavko nekega prejšnjega obračuna. Zato takšnih parov sploh ni potrebno prenašati v podatkovno skladišče. Ker pa se stornacije lahko dogajajo za nazaj in to za več let, je potrebno pri polnjenju podatkovnega skladišča to upoštevati in bodisi posodabljati, bodisi izbrisati in ponovno napolniti in preračunati vse postavke obračunane energije za več let.

Kot rečeno, podatki posameznega zapisa iz tega vira predstavljajo obračunano energijo za merilno mesto, za določeno obdobje, največkrat obdobje enega meseca. Primerjati želimo obračunane količine z izmerjenimi, ki so v podatkovnem skladišču obdelane na dnevnem nivoju. Zato moramo tudi obračunane količine nekako razčleniti na dnevni nivo. To naredimo tako, da skušamo za vsak zapis iz vira, skupno količino energije razdeliti po posameznih dneh njegovega obdobja, s kar najbolje določenimi deleži. Privzamemo, da merilna mesta, ki so merjena s pametnimi števci vključenimi v sistem AMI, predstavljajo zelo reprezentativen vzorec porabe. Kadar je merilno mesto, za katerega želimo porazdeliti obračunano porabo po dneh gospodinjsko, vzamemo množico vseh gospodinjskih izmerjenih porab, kadar je poslovno, pa množico vseh poslovnih izmerjenih porab. Za vsak dan obračunskega obdobja nato pomnožimo obračunano količino obdobja z dnevno vsoto izmerjenih porab prej omenjene množice in to delimo z vsoto izmerjenih porab iste množice za celotno obračunsko obdobje. To naredimo za vse zapise obračunane porabe. Za vsak zapis iz obračunskih podatkov, ki se nanaša na obdobje denimo N-tih dni, dobimo N zapisov. Za merilna mesta, ki imajo v nekem obdobju več prekrivajočih se obračunov (poračunov, …), dobimo za vsakega od dni s prekrivanjem po več zapisov. Porazdeljene obračunske vrednosti za posamezno merilno mesto, za vsak dan posebej, seštejemo. Tako pridemo do po enega zapisa z obračunano dnevno porabljeno količino energije za vsak dan, za vsako merilno mesto (tudi tista, s podatki o 15-minutnih porabah).

Za merilna mesta, ki so vključena v sistem naprednega merjenja (s 15-minutnimi ali urnimi količinami), želimo v isto tabelo dejstev ob meri 'obračunana dnevna količina energije', dodati še mero 'izmerjena dnevna količina energije'. To veličino bomo pridobili iz sistema naprednega merjenja, vendar bomo raje kot podatke iz registrov, ki vsebujejo količine energije za vsakih 15 minut, uporabili registre, ki nosijo podatke o števčnem stanju (za vsako tarifo – VT/MT/ET – posebej). Isti registri se uporabijo tudi pri podajanju mesečnih količin za obračun. Razlog, da ne uporabimo 15-minutnih podatkov je v tem, da so tam podane količine energije, porabljene v tem intervalu. Če želimo izračunati količino energije v daljšem intervalu, lahko seštejemo količine vseh pripadajočih 15-minutnih intervalov. Vendar se lahko zgodi, da kakšna od meritev izpade ali pa je odčitek neveljaven (zaradi težav na samem števcu ali v komunikaciji). Vsota količin vseh razpoložljivih pripadajočih intervalov v tem primeru seveda ne da prave količine. Registri s števčnimi stanji se zabeležijo v sistemu s po eno vrednostjo dnevno (ob koncu dneva). Razlika med poznejšim in zgodnejšim stanjem števca predstavlja porabljeno količino energije v tem obdobju. Tudi če kakšno vmesno stanje števca izpade ali je neveljavno, je razlika količin dveh števčnih stanj še vedno pravilna količina energije, porabljene v obdobju med tema dvema

odčitkoma. Količino za mero 'izmerjena dnevna količina energije' torej pridobimo tako, da od števčnega stanja določenega dne, odštejemo prejšnje števčno stanje istega merilnega mesta. Za merilna mesta, ki niso opremljena s pametnim števcem, ki je vključen v sistem AMI, v to mero vpišemo vrednost 0.

Naša tabela dejstev F_EnergijaMM tako vsebuje dve meri: obračunano energijo in izmerjeno energijo. En zapis v tej tabeli pa predstavlja podatke za eno merilno mesto za en dan.

4.2.2 Dimenzija čas

Časovna dimenzija je ena od najpomembnejših in je prisotna praktično v vsakem podatkovnem skladišču. Vsebina tabel, ki predstavljajo časovno dimenzijo, je lahko za različna skladišča, zelo podobna. Pomembno je seveda, kakšno časovno ločljivost potrebujemo v naših analizah. Glede na časovno ločljivost podatkov v naših tabelah dejstev je naravno, da si pripravimo časovno dimenzijo, v kateri en zapis predstavlja en dan.

V našem primeru poleg običajnih podatkov, ki se pojavljajo v takšnih tabelah (kot so: leto, kvartal, mesec, teden, dan v tednu,…), dodamo nekaj specifičnih podatkov (kot npr. praznik, dan pred praznikom, dan po prazniku, sezona, …).

Časovno dimenzijo napolnimo le inicialno in njene vsebine s postopkom ETL, ki se izvaja redno, ne spreminjamo.

4.2.3 Dimenzija merilno mesto

Entiteta 'merilno mesto' je ena od temeljnih pri delovanju elektrodistribucijskega podjetja. Predstavlja ekvivalent pojmu 'odjemno mesto', le da je izraz splošnejši, saj ni dileme, da vključuje tudi prejemna mesta proizvodnih objektov (kar pri 'odjemnem mestu' ni samoumevno). Najbolj elementarni način spremljanja porabe energije je po merilnem mestu, zaradi česar smo podatke o merilnih mestih dodali v podatkovno skladišče kot dimenzijo. Vir za podatke o merilnih mestih je sistem eIS, kjer nastajajo skozi procese življenjskega cikla odjemalca.

Atributi merilnega mesta, ki smo jih zapisali v podatkovno skladišče, se nanašajo na njegovo lokacijo, mesto priključitve v omrežju, pripadnost organizacijski enoti, ki je odgovorna za vzdrževanje, pripadnost odbirnemu okolišu in podobno. Vse te lastnosti so precej statične. Zato smo se odločili, da bo dimenzija merilno mesto časovnega tipa 1. To pomeni, da bomo, če se merilnemu mestu spremeni katera od lastnosti v transakcijskem sistemu, temu vsebino prepisali z novo vsebino in ne bomo sledili zgodovini sprememb. Analize, izvedene z atributi merilnega mesta po spremembi, bodo vrnile rezultate, kot če bi to merilno mesto imelo vedno takšne lastnosti.

Na vsakem merilnem mestu, preko katerega odjemalec aktivno prejema ali oddaja energijo, mora biti nameščen števec. Istočasno sme biti na merilnem mestu nameščena le ena števčna garnitura. Torej mora biti na vsakem aktivnem merilnem mestu vedno nameščen natanko en

števec. Podatek o tem, kateri števec je nameščen na nekem merilnem mestu, oz. obratno, na katerem merilnem mestu je nameščen nek števec, je zapisan v evidenci števcev.

Enako velja za pogodbo o dostopu. Vsako aktivno merilno mesto mora imeti natanko eno veljavno pogodbo o dostopu, po kateri se obračunava omrežnina. Podatek o tem, katera pogodba se nanaša na neko merilno mesto, oz. obratno, na katero merilno mesto se nanaša neka pogodba, je zapisano v evidenci pogodb o dostopu v sistemu eIS.

4.2.4 Dimenzija števec

Evidenca števcev se vodi v sistemu eIS. Isti števec je lahko skozi čas nameščen na različna merilna mesta. Ob osnovnih lastnostih števca se v eIS vodi tudi evidenca namestitev števcev tako, da je naveden datum vgradnje in datum odstranitve števca na določeno merilno mesto.

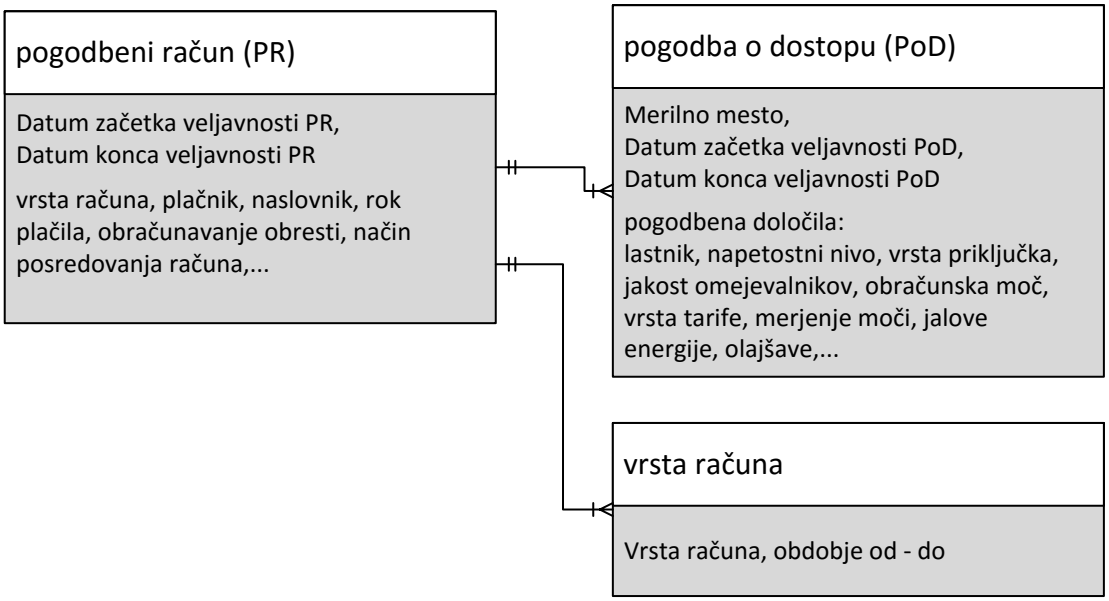
V podatkovnem skladišču bomo potrebovali dimenzijsko tabelo, ki opisuje števce, predvsem v povezavi s podatki iz sistema AMI. H kateremu merilnemu mestu pripada nek niz meritev, bomo povezali preko števca in podatka o njegovi namestitvi. Med atributi števca v dimenzijski tabeli zato potrebujemo tudi številko merilnega mesta in zmožnost sledenja spremembam vgradnje. Zaradi tega dimenzijo načrtujemo kot časovno spremenljivo, tipa 2. To pomeni, da se bo isti števec ob spremembi lastnosti (kot je npr. mesto vgradnje), pojavil v dimenzijski tabeli kot nov zapis. V prejšnjem zapisu se vpiše datum konca veljavnosti, pri novem zapisu, ki govori o trenutnem mestu namestitve, pa pustimo polje 'datum konca' prazno ali pa vpišemo datum z maksimalno možno vrednostjo. Ker je zgodovina namestitev na enak način zapisana tudi v izvorni tabeli transakcijskega sistema, nam 'rekonstrukcija zgodovine' ob polnjenju dimenzijske tabele ne dela nobenih težav.

Na težavo pa naletimo pri osnovnem – ključih za povezovanje. V sistemu AMI je števec določen le s tovarniško številko. V izvornem transakcijskem sistemu (eIS) je primarni ključ tabele, ki opisuje števce, sestavljen iz tovarniške številke in tipa števca. Razlog za takšno rešitev je, da so se pri različnih proizvajalcih in različnih tipih števcev njihove tovarniške številke podvajale. Res, da je to veljalo za starejše števce, a nekateri takšni so še v uporabi. Težava za enolično povezovanje števcev med eIS in AMI pa ostaja predvsem pri podatkih za preteklost. Ob izdelavi dimenzijske tabele števcev smo se skušali izogniti tej težavi tako, da vanjo izmed števcev, katerim se tovarniške številke podvajajo, prenašamo le po enega. Izločimo tiste, ki so bili prej odstranjeni iz uporabe (imajo nižji datum odstranitve). V kolikor imata dva števca z enako tovarniško številko tudi enak zadnji datum odstranitve, izločimo tistega z nižjo številko tipa števca. Na tak način res pridobimo vedno le zapise za en števec (pri tovarniških številkah, ki se ponavljajo za različne tipe števcev, upoštevamo le enega od tipov), vendar ni nujno, da je to vedno pravilno. Tak ukrep ne odpravlja težave, jo le delno zaobide.

4.2.5 Dimenzija pogodba o dostopu

Pogodba o dostopu v obračunskem sistemu eIS določa parametre za obračun. Pogodba o dostopu se vedno nanaša natanko na eno merilno mesto in ima določen datum začetka veljavnosti in datum konca veljavnosti. Skozi pogodbena določila pogodba določa parametre

kot so: lastnik, napetostni nivo, vrsta priključka, jakost omejevalnikov, obračunska moč, vrsta tarife, merjenje moči, jalove energije, olajšave, …



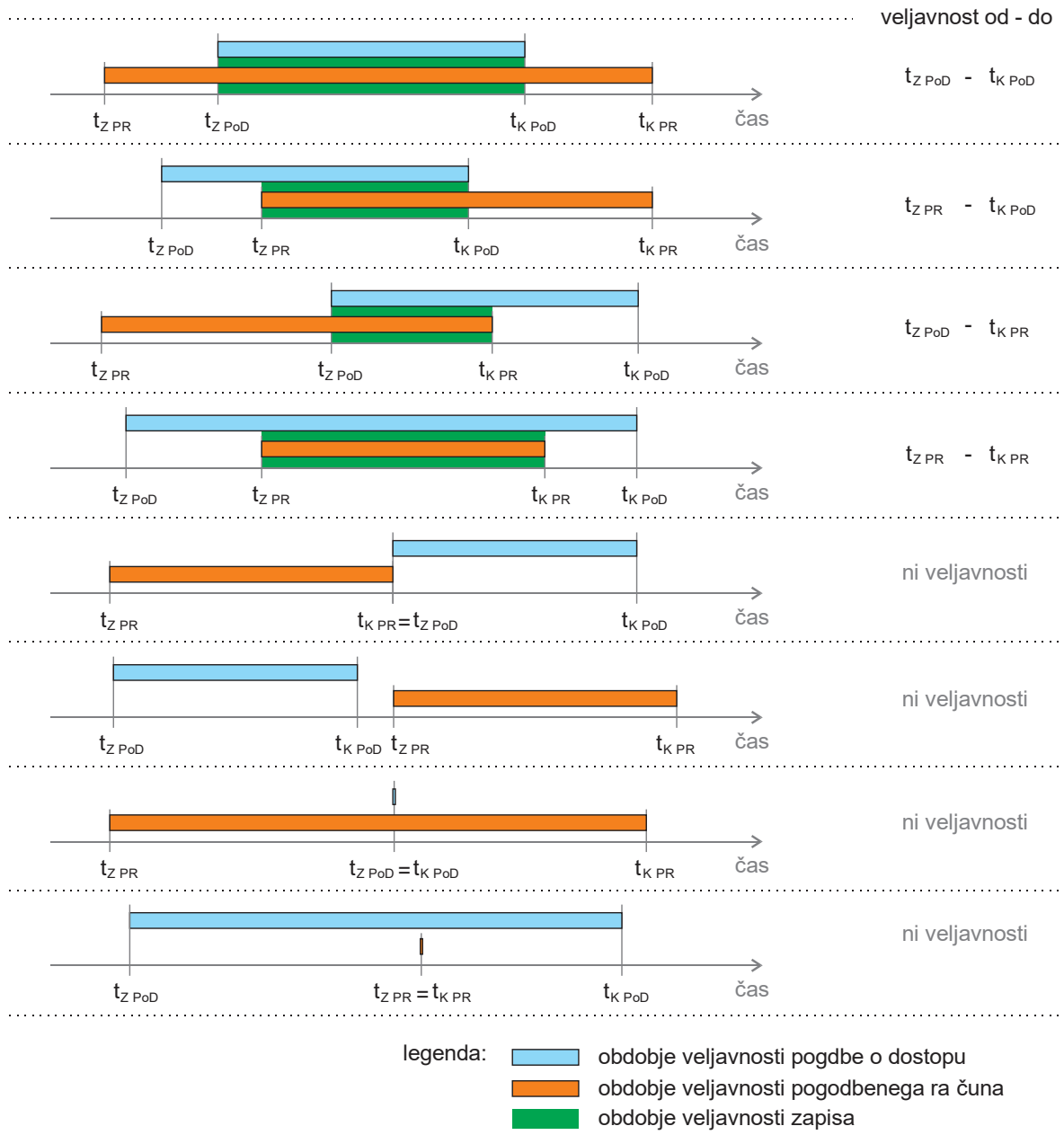
Slika 4.4: Poenostavljen entitetni diagram konceptov, ki jih združene (denormalizirano) zapišemo v dimenzijo 'pogodba o dostopu'

Pogodba o dostopu mora biti vključena na nek pogodbeni račun. Pogodbeni račun določa lastnosti kot so: plačnik, naslovnik, rok plačila, obračunavanje obresti, način posredovanja računa, … Na enem pogodbenem računu je lahko navedena ena ali več pogodb o dostopu, za katere se izdaja združen račun. Če jih je več, imajo vse skupne lastnosti, ki jih določa pogodbeni račun. Tudi pogodbeni račun ima določen datum začetka in datum konca veljavnosti.

Na pogodbeni račun se navezuje podrejena tabela, z zapisi za vrsto računa (skupen / ločen račun). Med obdobjem veljavnosti pogodbenega računa se lahko vrsta računa zanj večkrat spremeni. Zato ima vsak zapis v podrejeni tabeli vrst računa za pogodbeni račun določeno tudi obdobje veljavnosti tega podatka.

Eden od obveznih pogojev, da bo za neko merilno mesto izdelan obračun je, da ima to merilno mesto veljavno pogodbo o dostopu. Pogodbe o dostopu, ki se navezujejo na en pogodbeni račun, lahko imajo različna obdobja veljavnosti, ki se z obdobjem veljavnosti pogodbenega računa lahko pokrivajo delno, v celoti, ali pa celo segajo čez njegovo veljavnost.

Vir podatkov o pogodbi o dostopu in pogodbenem računu za nas bo obstoječ vpogled (view) v podatkovni bazi sistema eIS, v katerem so v posameznem zapisu združeni podatki o vseh treh prej omenjenih entitetah, prikazanih na sliki 4.4. Zapis vsebuje datuma začetka in konca veljavnosti pogodbe o dostopu, prav tako pa tudi datuma začetka in konca veljavnosti pogodbenega računa. Med zapisi, ki predstavljajo celotno zgodovino spreminjanja pogodb o dostopu, najdemo zapise z različnimi kombinacijami obdobij veljavnosti pogodbe o dostopu in pogodbenega računa, kot jih prikazuje slika 4.5.



Slika 4.5: Obdobje veljavnosti zapisa o pogodbi o dostopu

V dimenzijo 'Pogodba o dostopu' v podatkovnem skladišču smo tako prenesli denormalizirane podatke o pogodbi, razširjene s podatki o pogodbenem računu in vrsti računa. Pri tem smo v različnih kombinacijah veljavnosti le teh, kot obdobje veljavnosti zapisa, upoštevali zeleno označeno obdobje s slike 4.5. Zapise, ki nimajo obdobja veljavnosti, ker se obdobji ne prekrivata ali ker ima eno od obdobij 0 dni trajanja (kot je prikazano na sliki 4.5) ter zapise, ki imajo označen status pogodbe ali status pogodbenega računa z 'neveljavno', smo ignorirali.

Nekatere od teh situacij so povsem regularne. Gre namreč za napačne vnose, ki so bili stornirani in nadomeščeni z drugimi, veljavnimi.

4.2.6 Dimenzija omrežje

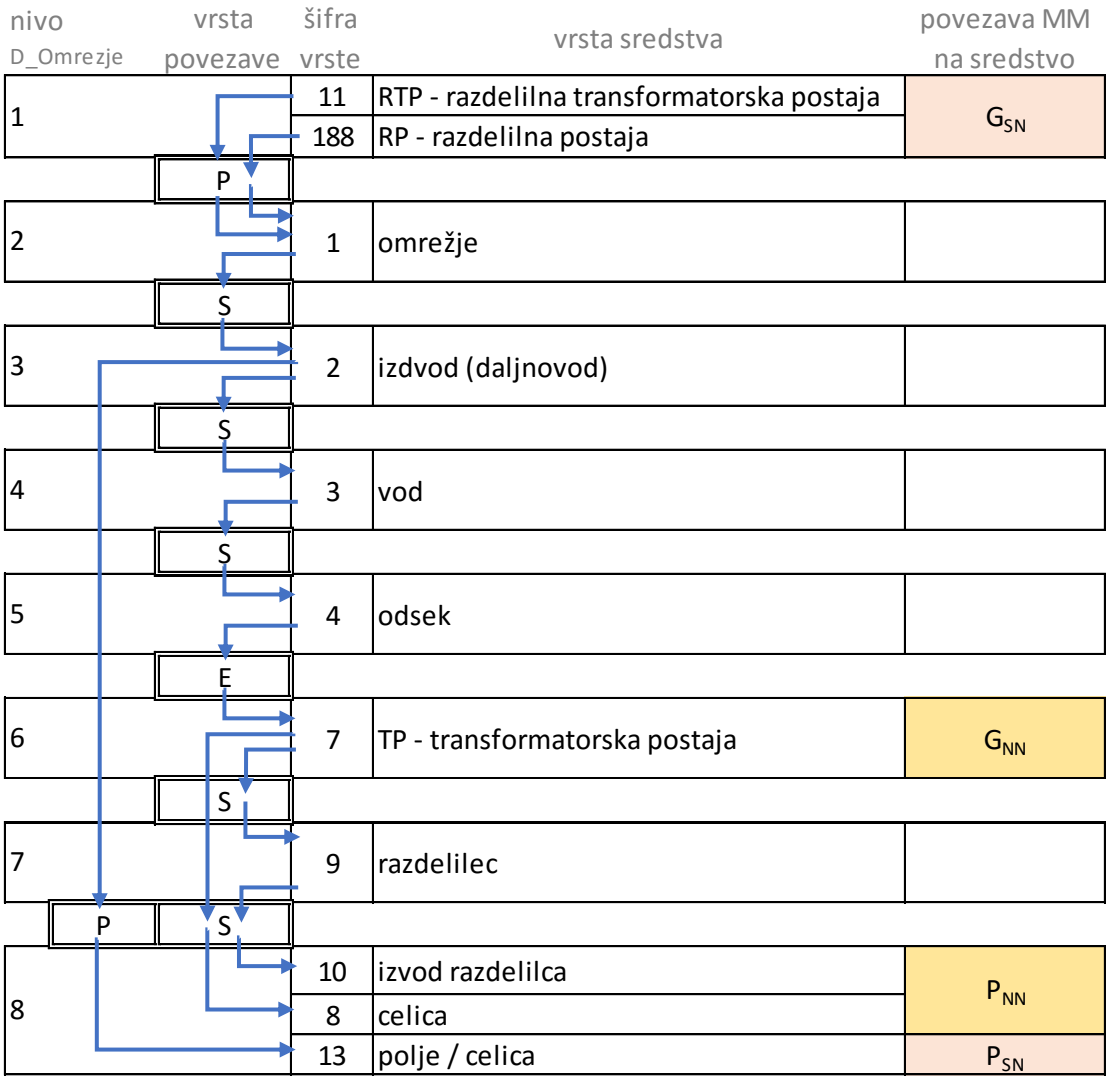
Elektrodistribucijsko omrežje je dokaj kompleksen sistem, v katerem je vključenih veliko sredstev različnih vrst. V Elektru Celje podatke o sredstvih elektrodistribucijskega sistema opišejo prvenstveno v prostorskem informacijskem sistemu (GIS). Vnesejo lastnosti sredstev in njihovo strukturo (povezave med sredstvi). Del teh podatkov prepisujejo (sinhronizirajo) v starejši zaledni sistem, kjer se struktura, ki opisuje elemente in sestavo omrežja, imenuje BTP - 'baza tehničnih podatkov'. Zaradi zahteve po tej povezanosti informacijskih sistemov mora tudi sistem GIS upoštevati nabor vrst sredstev in pravila za njihovo povezovanje, kakor jih določa BTP.

Pri analizi porabe električne energije po merilnih mestih, nas bo zanimal le manjši del te vsebine in sicer tisti del, ki govori o povezovanju omrežja od merilnih mest 'navzgor' do razdelilnih transformatorskih postaj, kjer distribucijsko podjetje prevzame energijo iz visokonapetostnega prenosnega omrežja. Za potrebe združevanja količin z najnižjih (najpodrobnejših) nivojev, tj. od merilnih mest, proti najvišjemu (najsplošnejšemu) nivoju, tj. k razdelilnim transformatorskim postajam, smo si v dimenzijski tabeli, ki opisuje omrežje, zamislili členitev na 8 nivojev. Po teh nivojih bomo nato lahko izdelali hierarhijo, ki jo bomo uporabili pri analizah. Slika 4.6 prikazuje pravila, ki jih upoštevamo pri transformaciji opisa omrežja iz strukture BTP v dimenzijsko tabelo, ki opisuje omrežje v podatkovnem skladišču. Osem nivojev, ki jih slika navaja, je predstavljenih z osmimi pari atributov, v katere vpišemo šifro in naziv sredstva. En zapis v dimenzijski tabeli tako dejansko predstavlja sredstvo najnižjega (osmega) nivoja, ki pa ima v vrednostih atributov višjih nivojev opisano njihovo povezavo vse do razdelilne transformatorske postaje. Vrste sredstev, ki jih vpišemo na najnižjem nivoju so lahko:

- izvod razdelilca - nizkonapetostni izvod transformatorske postaje
- celica – nizkonapetostna celica v transformatorski postaji
- polje/celica – srednjenapetostno polje ali celica v razdelilni postaji ali v razdelilni transformatorski postaji

Za ta nivo podrobnosti smo se odločili zato, ker v sistemu eIS merilnim mestom določimo mesto priključitve na omrežje tako, da atributu 'sredstvo_podrejeni_element' priredimo šifro sredstva vrste 'izvod razdelilca' ali 'celica', če gre za merilno mesto na nizki napetosti in šifro sredstva vrste 'polje/celica' če gre za merilno mesto na srednji napetosti. Merilno mesto pa ima poleg tega še en atribut, ki govori o mestu njegove priključitve na omrežje. To je atribut 'sredstvo_glavni_element'. Načeloma bi povsem zadoščalo, če bi bilo mesto priključitve vedno opredeljeno le na najnižjem nivoju, torej s 'sredstvo_podrejeni_element', saj lahko nadrejene nivoje po strukturi, opisani na sliki 4.6 vedno določimo. V primeru, kot ga imamo sedaj, pa 'sredstvo_glavni_element' predstavlja redundantno informacijo in daje možnost za vnos napak in s tem za zmešnjavo. Razlogi, da obstajata dva podobna atributa merilnega mesta, ki podajata mesto njegove priključitve na omrežje so zgodovinski. Sprva je obstajal le 'glavni_element'. Kasneje, ko se je izkazalo, da je potrebno mesto priključitve opredeliti podrobneje, je bil dodan še 'podrejeni'. Vendar pa vsebina novega atributa ob uvedbi te spremembe ni bila določena za

vsa merilna mesta, ampak so jo uporabniki izpolnjevali pozneje. Tudi še danes je precejšen delež merilnih mest, ki nimajo vpisane vrednosti za ta atribut.



Slika 4.6: Preslikava omrežja iz BTP v dimenzijo 'omrežje' (sredstva različnih vrst, upoštevaje navedene vrste medsebojnih povezav, zapišemo v dimenzijsko tabelo po nivojih)

Za določitev povezave merilnih mest (preko atributa 'sredstvo_podrejeni_element') na omrežje, bomo zato v našem podatkovnem skladišču uporabili sredstvo osmega nivoja (za vsa merilna mesta; na nizki in srednji napetosti). Merilna mesta, ki jih na tak način ne bomo uspešno 'povezali v omrežje', bomo skušali povezati preko atributa 'sredstvo_glavni_element' za nizkonapetostna sredstva šestega nivoja, za srednjenapetostna pa sredstva prvega nivoja.

Ob pripravi postopka za polnjenje dimenzije 'omrežje' in analizi izvornih podatkov smo ugotovili, da se srečamo s kar precej anomalijami v podatkih. Tako obstajajo npr. transformatorske postaje, ki nimajo opredeljene povezave tipa 'E' z odsekom ali pa razdelilne transformatorske postaje ali razdelilne postaje, ki niso povezane z nobenim omrežjem ali da obstajajo vodi brez odsekov ali transformatorske postaje brez izvodov in brez celic in podobno. Da ne bi iz dimenzijske tabele izpadli vsi izvodi in celice ter transformatorske postaje, ki niso

pravilno povezani v celovito strukturo, smo v takšnih zapisih v višje atribute vpisali, da se npr. takšna transformatorska postaja nahaja na nedoločenem odseku, na nedoločenem vodu, … na nedoločeni razdelilni transformatorski postaji. Zaradi kasnejšega povezovanja merilnih mest, ki nimajo izpolnjenega atributa 'sredstvo_podrejeni_element', smo dodali še po en zapis za vsako transformatorsko postajo in zanjo vpisali 'nedoločen razdelilec' in 'nedoločen izvod razdelilca' oz. 'nedoločeno celico'.

Spremembe v omrežju se dogajajo dokaj pogosto. Zgradijo se novi deli omrežja, ponekod se stari deli razgradijo, spremeni se topologija, obstoječi odseki se razbijejo v več novih in podobno. V izvornih podatkih se to odraža tako, da imajo sredstva različne statuse (npr. 'v obratovanju' ali 'izbrisano oz. odstranjeno' ali 'rezerva'…) in da imajo povezave med sredstvi poleg že omenjenega tipa povezave določen tudi datum začetka veljavnosti in datum konca veljavnosti relacije. Ob razmisleku ali je to dinamiko smiselno upoštevati v podatkovnem skladišču ali ne, smo se odločili, da ne, saj bi ta dodatna kompleksnost (ki je odraz realnosti) verjetno predstavljala tudi težavo za uporabnike pri interpretaciji rezultatov. Pri vsaki izvedbi postopka posodabljanja vsebine dimenzije 'omrežje' pridobimo stanje celotnega omrežja in to primerjamo s prejšnjim stanjem. Naravni ključ za povezovanje teh vsebin je za zapise, ki imajo izpolnjeno šifro transformatorske postaje (šifro sredstva šestega nivoja), sestavljen iz šifer sredstev prvega, šestega, sedmega in osmega nivoja. Za zapise, ki imajo na 6. nivoju vpisano 'neobstoječe sredstvo', moramo vsebine povezati preko šifer sredstev vseh ostalih nivojev. Vse zapise, ki se na tak način povežejo, v celoti posodobimo (prejšnje vsebine prepišemo z novimi). Zapise, ki so v novem stanju, v starem pa jih ni, dodamo. Zapise, ki so v starem stanju, preko prej omenjene povezave pa jih ne najdemo v novem stanju, v skladišču označimo z 'neaktualno' (jih ne brišemo). S takšno vrsto časovne spremenljivosti dimenzije (tip 1), dobimo rezultate analiz po spremembi omrežja takšne, kot če bi omrežje imelo takšno strukturo vedno v preteklosti.

V vsakem primeru s podatki o topologiji omrežja iz opisanega vira seštevanje energije 'od spodaj navzgor', proti razdelilnim transformatorskim postajam, nikoli ne bo moglo dati absolutno pravilnih rezultatov za daljše obdobje ker podatki o topologiji iz GIS/BTP predstavljajo le privzeto topologijo. Vendar pa v omrežju obstajajo tudi zanke in povezovalni prečni vodi, preko katerih je s preklopom progovnih stikal možno spremeniti topologijo. Te spremembe se izvajajo dinamično, bodisi zaradi izvajanja del v omrežju, bodisi zaradi zagotavljanja višje razpoložljivosti, bodisi zaradi optimizacije izgub. V teh primerih seštevanje energij po privzeti topologiji seveda ne odraža dejanskega stanja. V kolikor bi imeli na voljo podatke o trenutni topologiji omrežja, bi bilo vredno ponovno razmisliti o predelavi dimenzije 'omrežje' v časovno spremenljivo, tipa 2 (z beleženjem zgodovine stanj).

4.2.7 Dimenzija transformatorska postaja

Zelo sorodna dimenziji 'omrežje' je dimenzija 'transformatorska postaja'. Zapis v tej dimenzijski tabeli predstavlja en razdelilec v transformatorski postaji. V nasprotju z dimenzijo 'omrežje' ta ne vsebuje podatkov o povezovanju v omrežju. Vir za podatke o transformatorskih postajah in

njihovih razdelilcih je sistem GIS. Podatki, ki jih pri tem pridobimo, so: tip, status, vrsta (mestna/podeželska), lastnik, organizacijska enota, zadolžena za vzdrževanje, število transformatorjev, njihova moč in nenazadnje, tovarniška številka pametnega števca, ki je vgrajen v transformatorski postaji.

V splošnem bi lahko s podatki, ki jih vpišemo v dimenzijo 'transformatorska postaja', enostavno razširili dimenzijo 'omrežje'. Za to se nismo odločili, ker je med transformatorskimi postajami tudi nekaj takšnih, ki niso v lasti Elektra Celje in bi zato imeli težave z njihovo navezavo na omrežno strukturo. Vsekakor bi tudi to lahko rešili z navezovanjem na 'nedoločena sredstva' in združevanjem dveh dimenzij v eno. Vendar pa smo dimenzijo 'transformatorska postaja' povezali z dimenzijo 'omrežje' tako, da smo vsakemu zapisu, kateremu lahko identificiramo istopomenski zapis v le tej, pripisali šifro tega zapisa.

4.2.8 Avtomatizacija procesa polnjenja skladišča – postopek ETL

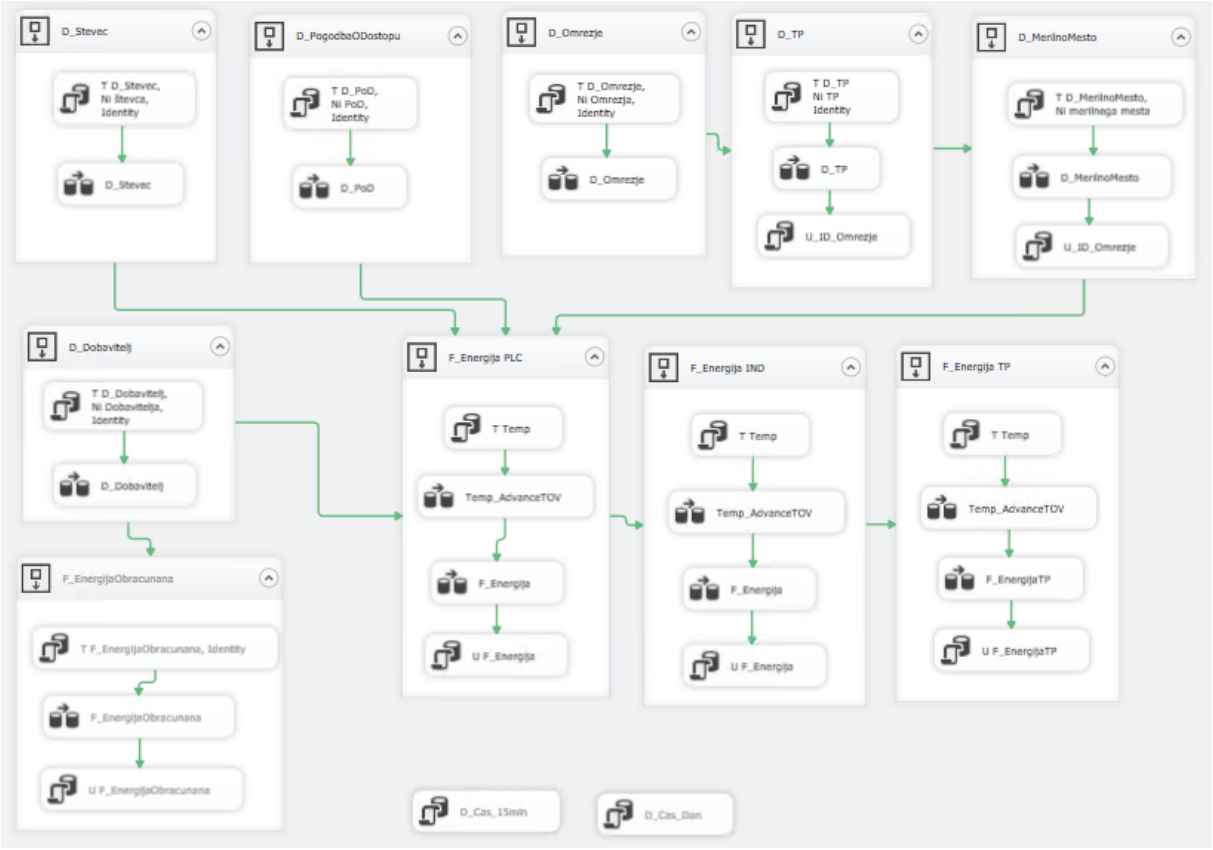
Priprava podatkov v podatkovnem skladišču je tisti del projekta poslovne inteligence, ki poslovnih uporabnikov običajno ne zanima pretirano. Pričakujejo le, da bodo podatki za analizo pripravljeni čim prej, da bodo konsistentni, vedno ažurni in da bo možno strukturo sproti prilagajati trenutnim potrebam. Realnost žal pogosto kaže ravno nasprotno sliko. Postopki za zajem, preoblikovanje in polnjenje podatkov, ki jih pojmujemo pod uveljavljeno oznako ETL (angl. extract, transformation and load), so običajno kompleksni. Zaradi tega načrtovanje strukture in izgradnja podatkovnega skladišča običajno traja veliko dlje, kot so uporabniki pripravljeni razumeti. Kljub vsemu pa tukaj ne gre iskati bližnjic, saj lahko pomeni na videz zelo 'nedolžna' zahteva po naknadni razširitvi skladišča (npr. z dodatnim nivojem podrobnosti), da je potrebno tako skladišče, kot postopek za njegovo polnjenje in vzdrževanje, v celoti ponovno izdelati. Če je bila prvotna zasnova slaba, je verjetnost, da se nam bo to zgodilo, še toliko večja. Poleg tega lahko slaba zasnova pripelje do nepreglednosti in do tega, da so rezultati analiz povsem drugačni, kot bi pričakovali. Kadar imamo opravka z velikimi količinami spremenljivih podatkov, je običajno izziv tudi to, kako optimirati izvajanje procesa ETL za redno vzdrževanje podatkov v skladišču, da bodo uporabniku res na voljo ažurni takrat, ko jih potrebuje.

Platforma Microsoft SQL Server, na kratko opisana v točki 4.1, nam v ta namen ponuja SQL integracijske storitve SSIS (angl. SQL Server Integration Services) in razvojno orodje SSDT (angl. SQL Server Data Tools). V tem okolju zlahka vzpostavimo (in pozneje upravljamo) povezave z različnimi vrstami virov strukturiranih in nestrukturiranih podatkov. V našem primeru imamo tri izvorne sisteme: AMI, eIS in GIS. Vsi trije hranijo podatke v relacijskih bazah in nam omogočajo dostop do njih tudi preko pripravljenih pogledov (angl. view) na nivoju mehanizma relacijske baze. To možnost smo tudi izkoristili, saj zagotavlja velike prepustnosti in hitrosti. Povezave do podatkovnih baz izvornih sistemov smo tako vzpostavili preko ponudnikov OLE DB (angl. OLE DB provider).

V razvojnem okolju SSDT gradimo proces ETL z grafičnim uporabniškim vmesnikom z gradniki, med katerimi s povezavami v obliki puščic določimo tok podatkov in izvajanja.

Gradniki nam ponujajo bogato paleto možnosti za preoblikovanje podatkov. Postopek, ki ga sestavimo s temi gradniki, shranimo kot paket (angl. SSIS package). Če želimo izvajanje postopka avtomatizirati, lahko paket objavimo v repozitoriju integracijskih storitev (SSISDB) in zanj nastavimo urnik izvajanja ali izvajanje vežemo na določen dogodek oz. klic. Izvajanje postopka (oz. paketa) lahko prožimo kar v razvojnem okolju SSDT (v celoti ali le posameznih gradnikov oz. sklopov - kar je priročno v času razvoja), ne da bi ga objavljali v repozitoriju.

Ob prvem polnjenju podatkovnega skladišča moramo napolniti tudi vsebine, ki se pozneje ne bodo spreminjale ali se bodo spreminjale zelo redko, le na zahtevo. Tipično je takšna časovna dimenzija. Pri večini dimenzij in dejstvih se bomo ob rednih osveževanjih skušali omejiti na prenašanje le spremenjenih in dodanih podatkov. Zato bodo pri večini dimenzij postopki pridobivanja, ponekod pa tudi transformacije (kadar bo npr. potrebno rekonstruirati zgodovino sprememb za dimenzijsko tabelo s časovno spremenljivostjo tipa 2) za prvo polnjenje drugačne, kot za naslednja. Čeprav prvo polnjenje načeloma izvedemo le enkrat, je smiselno posamezne korake postopka sestaviti v proceduro. Proceduro, s katero smo izvedli začetno polnjenje našega podatkovnega skladišča, prikazuje slika 4.7.



Slika 4.7: ETL postopek za inicialno polnjenje podatkovnega skladišča

Kot vidimo na sliki, najprej napolnimo dimenzijske tabele, da lahko nato pri polnjenju tabel dejstev zapise dejstev povezujemo z ustreznimi zapisi v dimenzijskih tabelah. Kot velja splošno pravilo oz. dobra praksa za izgradnjo podatkovnih skladišč, vsem dimenzijskim tabelam

določimo neodvisne lastne primarne ključe v obliki celoštevilčnih števil, ki se ob dodajanju zapisov samodejno številčijo. Vendar še preden vključimo samoštevilčenje, v tabelo dodamo zapis (običajno z vrednostjo ključa – identifikacijske številke 0), ki ga bomo povezali z zapisom dejstva, kadar bomo naleteli na nedoločenost ali nekonsistentnost. Procedura ETL za začetno polnjenje vsebuje tudi ta del postopka.

Slika 4.8: ETL postopek za redno (dnevno) dodajanje in posodabljanje podatkovnega skladišča

Procedura ETL za redno vzdrževanje vsebin v podatkovnem skladišču je nekoliko drugačna. Prikazana je na sliki 4.8. Glede na naravo podatkov iz sistema AMI, ki so osrednja vsebina,

okoli katere gradimo skladišče, je smiselno razmišljati o izvajanju posodobitev enkrat dnevno – usklajeno z dnevnim zbiranjem podatkov s pametnih števcev v sistemu AMI.

Tukaj zapise v dimenzijskih tabelah le dodajamo in posodabljamo. Pri tem si pomagamo z začasnimi vmesnimi tabelami, kar se je izkazalo za izjemno učinkovito (hitro) tudi v primeru velikega števila zapisov. Ko ugotavljamo, kateri izvorni zapisi so se spremenili od prejšnjega izvajanja posodabljanj, si lahko pomagamo s časovnimi značkami sprememb zapisov iz izvornih sistemov, če ti obstajajo. V nekaterih primerih naših vhodnih podatkov takšen atribut sicer obstaja, a smo skozi analizo ugotovili, da se nanj ne moremo zanesti. Naleteli smo namreč na primere, ko je bila vsebina zapisa spremenjena, časovna značka pa ni kazala na to. Verjetno je razlog v tem, da pogledi v izvornem sistemu, ki združujejo podatke iz več elementarnih tabel, ne realizirajo korektno podatka o času zadnje spremembe (ne upoštevajo spremembe kjerkoli v celotnem sestavljenem zapisu). Zato smo se glede na naravo možnih sprememb vsebin odločali, za kakšno obdobje v preteklost bomo vsakič pridobili podatke, da jih bomo primerjali z vsebino v skladišču. Kadar je narava vhodnih podatkov takšna, da se lahko spremeni katerikoli podatek, vsakič prenesemo celotno vsebino in jo primerjamo s ciljno. Slika 4.8 prikazuje postopek ETL, namenjen rednemu dnevnemu posodabljanju podatkovnega skladišča.

Tudi pri vsebinah za tabele dejstev se lahko zgodi, da je potrebno posodobiti ali dodati zapise za obdobje, ki smo ga v prejšnjih posodabljanjih že obdelali. Tak scenarij je povsem realen pri podatkih s pametnih števcev. Kadar se zgodi, da zaradi komunikacijskih težav sistem AMI ni mogel pridobiti meritev določenega števca, bo pri naslednjem uspešnem pozivanju (npr. čez en dan) pridobil manjkajoče meritve, saj števec hrani vsebine svojih registrov za več dni (tipično za 40 dni). Ker v bazi AMI nimamo informacije o tem, kdaj je bil kateri od števčnih podatkov pridobljen v osrednjo bazo, je 'najcenejša' možnost, ki nam preostane ta, da pri vsakem posodabljanju izbrišemo vse podatke za zadnjih nekaj dni in jih skupaj s spremembami ponovno napolnimo. Običajno, kadar ne zaznavamo težav s komunikacijami do števcev, je lahko to obdobje zelo kratko (npr. dva dni), v primeru zaznanih težav, pa ga lahko povečamo. Da lahko to storimo brez spreminjanja kode v paketu, si število dni posodabljanja za določeno vrsto podatkov zapišemo v spremenljivko, kot parameter SSIS paketa.

4.2.9 Kakovost podatkov

Že v predhodnih točkah (4.2.4, 4.2.5 in 4.2.6) smo opisali okoliščine in vzroke v izvornih sistemih, ki pri posameznih dimenzijah povzročajo nekonsistentnosti. Kakovost podatkov in zagotavljanje absolutne konsistentnosti je izjemno pomembno tako za pravilnost analiz, kot seveda tudi za pravilno izvajanje poslovnega procesa, v našem primeru za pravilno obračunavanje omrežnine. V procesu preprečevanja odtekanja prihodkov zato veliko pozornosti namenjamo kakovosti podatkov.

Za zaznane težave predlagamo naslednje rešitve:

4.2.9.1 Težava pri povezovanju števcev iz sistema AMI z eIS preko tovarniške številke

Posledice težave s potencialno nekonsistentnim povezovanjem identitete števca med sistemoma AMI in eIS, opisane v točki 4.2.4 so lahko tudi takšne, da v obračunski sistem (eIS) iz sistema AMI dobimo napačne vhodne podatke in izdelamo napačen obračun. Težava, ki smo jo ob izgradnji podatkovnega skladišča zaznali, je takšne vrste, da se lahko pojavi še pri dodatnih primerih in redno povzroča napake. Če na kratko ponovimo: tovarniška številka števca, s katero v sistemu AMI določajo identiteto števca in ki služi za povezavo s sistemom eIS, sama po sebi ne zadošča, saj se v celotni populaciji števcev pojavijo sicer redki primeri različnih števcev, z enako tovarniško številko. Za enolično identifikacijo števca, je potrebno poleg tovarniške številke navesti še tip števca.

Težavo je potrebno odpraviti na viru, saj bodo sicer možne napake tudi v prihodnje. Sistema eIS in AMI morata pri izmenjavi identificirati števce z enovito oznako – ključem, ki jih enolično določa. To lahko dosežemo tako, da pri oznaki števca v sistemu AMI, pred tovarniško številko števca dodamo šifro tipa števca. To je potrebno narediti za vse števce. Nato je potrebno spremeniti mehanizme prenašanja podatkov iz sistema AMI v eIS in upoštevati nov ključ za povezovanje.

Druga možnost rešitve, ki je le pogojna, je ta, da v naravi izločimo iz uporabe vse starejše števce, katerih tovarniška številka se podvaja z drugim, novejšim števcem. V tem primeru bi vsaj za v bodoče zagotovili, da ne bi več prišlo do podvajanj tovarniških številk in pri prenosu podatkov iz AMI teh na bi mogli pripisati drugemu števcu in obračunati drugemu odjemalcu.

4.2.9.2 Težave z evidenco namestitev števcev v eIS

Ob pripravi postopka za polnjenje tabele dejstev z izmerjeno energijo, smo ob povezovanju na dimenzijo števec ugotovili, da v evidenci namestitev števcev (eIS) obstajajo zapisi o posameznih števcih, ki so bili istočasno nameščeni na več kot enem merilnem mestu. Tovrstnih zapisov je relativno malo, pa še ti so večinoma več let stari. Vendar je zaskrbljujoče to, da se pojavi en tovrsten primer za namestitev števca, ki je bila opravljena po uvedbi novega obračunskega sistema eIS, kar kaže na to, da je takšno napako na nek način možno ustvariti tudi v sistemu eIS.

Za celovito rešitev tega problema, bi bilo potrebno v testnem okolju sistema eIS opraviti dodatna testiranja in poskušati ponoviti napačen vnos. Ko bi tako ugotovili, pod kakšnimi pogoji je možno opraviti napačen vnos, bi lahko vgradili v sistem dodatno kontrolo, ki bi to onemogočala. V vsakem primeru, tudi če ne uspemo reproducirati napake, pa je smiselno izdelati poročilo, ki prikazuje vse števce iz trenutne evidence namestitev, ki imajo več kot eno aktivno namestitev. Takoj, ko se na takšnem poročilu pojavi kakšen zapis, je to znak za alarm, da je potrebno napako popraviti in raziskati, pod kakšnimi pogoji je lahko nastala.

Druga podobna napaka, ki smo jo zasledili na enak način kot prvo, je možnost, da je bilo po evidenci namestitev števcev sodeč, možno evidentirati namestitev števca tako, da sta bila na enem merilnem mestu hkrati aktivno nameščena dva števca. Zadnji tak primer zasledimo v letu

2012, to je pred uvedbo sistema eIS, kar navaja na misel, da v novem sistemu te napake ni možno ponoviti in da ni potrebno nič spreminjati.

Tudi za tovrstne primere je najbolje narediti poročilo, ki nas bo opozorilo v primeru, če se na kakšnem merilnem mestu hkrati pojavi več kot en števec.

Tretja možna napaka, vezana na namestitev števcev je situacija, ko na merilnem mestu, za katerega obstaja veljavna pogodba o dostopu, v evidenci namestitev števcev ni nobenega števca, vezanega na to merilno mesto. Tudi takšne napake lahko zasledujemo s pripravljenim poročilom, ki nas bo opozorilo, če se bo pojavilo kakšno aktivno merilno mesto brez evidentiranega števca.

4.2.9.3 Težave z anomalijami pri pogodbi o dostopu

V točki 4.2.6 smo opisali situacije z veljavnostjo pogodbe o dostopu in veljavnostjo pogodbenega računa, ter kako smo jih reševali pri pripravi mehanizma za polnjenje dimenzijske tabele.

Pri povezovanju zapisov v tabeli dejstev z dimenzijo 'pogodba o dostopu' smo zasledili še, da so v preteklosti obstajala merilna mesta, ki so imela več kot eno veljavno pogodbo o dostopu hkrati.

Odkrili pa smo tudi primere merilnih mest z nameščenim števcem in celo izmerjeno količino energije, ki za določeno, običajno zelo kratko obdobje niso imela veljavne pogodbe o dostopu. Če bi pri prvi vrsti napake pomenilo, da bodo iste količine zaračunane dvakrat, bi pri slednji energija sploh ne bila obračunana.

Rešitev, ki jo predlagamo, je podobna, kot pri točki 4.2.9.2. Pripravimo torej dve poročili: eno, ki prikazuje merilna mesta z več kot eno trenutno veljavno pogodbo o dostopu in drugo z merilnimi mesti, ki imajo nameščen števec in morebiti izmerjeno porabo, a nimajo nobene veljavne pogodbe o dostopu. Obe poročili nas morata opozoriti, če se na njih pojavi kakšen zapis. Ko se to prvič zgodi, moramo takoj odpraviti anomalijo, nato pa raziskati okoliščine, ki so omogočile njen nastanek. Z vgradnjo dodatnih kontrol in opozoril v sistem eIS je potrebno onemogočiti ponovni vnos takšnih napak.

4.2.9.4 Pomanjkljivosti in napake pri opisovanju omrežja

V točki 4.2.6 smo opisali strukturo, ki opisuje omrežje oziroma le tisti del omrežja, ki nam je zanimiv v kontekstu priključitve merilnih mest. Struktura opisuje topologijo omrežja od razdelilnih transformatorskih postaj, ki predstavljajo vhodne točke, do odjemnih mest, ki predstavljajo izhodne točke. Strukturo bomo uporabili za seštevanje porabe energije od izhodnih proti vhodnim točkam. Kot smo videli, je v strukturi zaslediti relativno veliko napak in še veliko več nepopolnosti. Predvsem nepopolnosti v smislu manjkajočih povezav med elementi omrežja. Zaradi teh hib ne bo prišlo do izpada obračuna ali napačnega obračuna omrežnine. Vendar nam takšno stanje otežuje analizo tako tehničnih kot komercialnih izgub.

Za izboljšanje stanja bi zagotovo koristila kakšna dodatna kontrola, vgrajena v sistem GIS, s katerim se izvaja vnos podatkov o omrežju.

Še bolj kot to pa je potrebno zgraditi zavedanje, da pravilno in celovito opisano omrežje lahko prinese mnoge koristi. S tem zavedanjem in močno podporo vodstva bo potrebno skozi projektni pristop opraviti čiščenje obstoječih in vnos manjkajočih podatkov. Pri tem lahko veliko pomagajo razne navzkrižne poizvedbe, ki opozarjajo na anomalije, dokler le te obstajajo. Hkrati z izvedbo projekta čiščenja in dopolnjevanja je potrebno zagotoviti, da bodo skozi redni proces evidentiranja sprememb v omrežju res pravilno in celovito evidentirane vse spremembe.

4.2.9.5 Pomanjkljivosti in napake pri določanju mesta priključitve merilnega mesta na omrežje

V točki 4.2.6 smo opisali zadrego, ki jo ima sistem eIS z določanjem točke priključitve merilnega mesta na omrežje s pomočjo dveh atributov: 'sredstvo_glavni_element' in 'sredstvo_podrejeni_element'. Tudi ta pomanjkljivost neposredno ne povzroča napak v obračunu. Vsekakor pa enako kot pomanjkljivosti v opisu omrežja (4.2.9.4) onemogoča kakovostnejšo analizo izgub z neposrednim primerjanjem izmerjenih količin. Pravilen podatek o mestu priključitve za vsa merilna mesta bi podjetju koristil še na mnogih področjih, ne 'le' pri preprečevanju odtekanja prihodkov.

Recept za ureditev stanja, ki ga predlagamo, je podoben kot pri točki 4.2.9.4 in sicer skozi projektni pristop uvodoma urediti stanje. Pri tem lahko zelo pomagajo poročila, ki prikazujejo pomanjkljivo in napačno povezana merilna mesta. Razlika v primerjavi s točko 4.2.9.4 je v tem, da je za določanje točke priključitve merilnega mesta na omrežje v sistem eIS nujno predhodno vgraditi kontrole, ki bodo enake na vseh mestih, kjer je možno ta podatek vnesti ali spremeniti. To pomeni, da bi morali preverjati, ali sredstvi, ki sta navedeni, ustrezata dovoljeni kombinaciji vrst sredstev ter ali med njima obstaja ustrezna povezava. Z ustrezno logiko, vgrajeno v uporabniški vmesnik, lahko preprečimo napake, kot jih opažamo sedaj oz. lahko poleg tega naredimo uporabniški vmesnik veliko učinkovitejši za uporabo. Ne moremo pa preprečiti, da bi uporabnik s tem vmesnikom izbral napačno točko priključitve merilnega mesta.

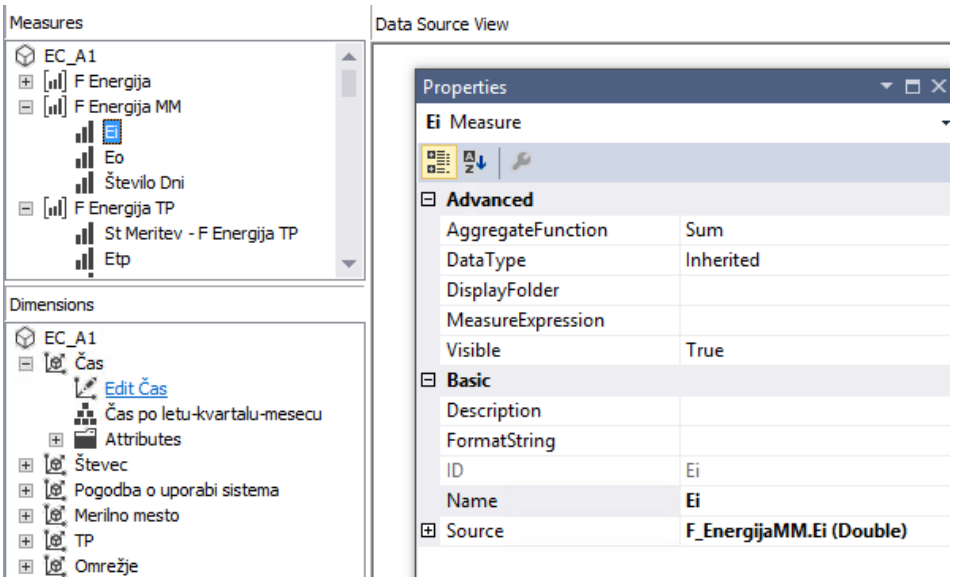
4.3 Analitična struktura (OLAP)

Kratica OLAP, ki pomeni sprotno analitično obdelavo (angl. Online analytical processing) je postala sinonim za programsko opremo, ki omogoča analizo velikih količin podatkov. Značilnost OLAP sistemov je, da podatke hranijo v posebni večdimenzionalni strukturi, imenovani OLAP kocka. Po vnaprej pripravljenem modelu dejstev in dimenzij, ki ta dejstva opisujejo, ob procesiranju mehanizem napolni OLAP kocko, v katero se zapišejo tudi že preračunani agregirani podatki.

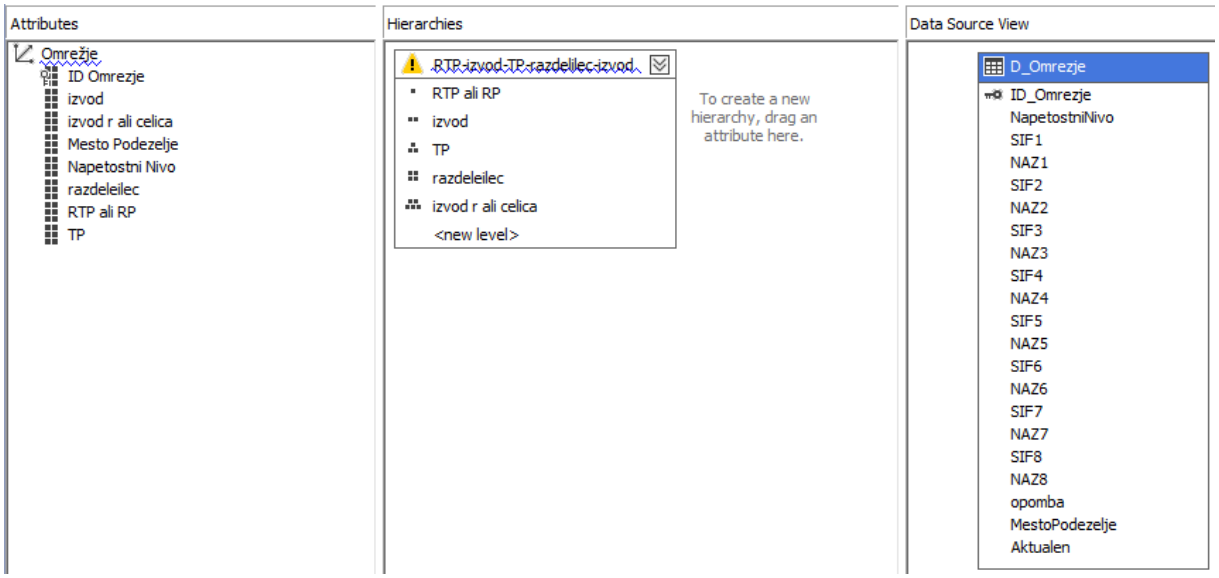
V infrastrukturi, ki smo jo izbrali za praktično izvedbo naše naloge in ki je opisana v točki 4.1, zmogljivost OLAP zagotavljajo analitične storitve SQL strežnika, s kratico poimenovane SSAS (angl. SQL Server Analysis Services). Model za to strukturo pripravimo v razvojnem okolju

SSDT (angl. SQL Server Data Tools). Izhodišče za pripravo analitične strukture je podatkovno skladišče, predstavljeno v točki 4.2.

Tabele dejstev predstavljajo skupine meritev. Ob načrtovanju strukture OLAP kocke izmed atributov tabele dejstev navedemo tiste, ki predstavljajo mere. Za vsako mero določimo tudi ustrezno agregacijsko funkcijo. Na sliki 4.9 vidimo uporabniški vmesnik okolja SSDT, s katerim določamo strukturo kocke. V razdelku levo zgoraj so navedene mere. Na desni strani iste slike vidimo lastnosti izbrane mere Ei iz skupine mer 'F Energija MM'. Vidimo, da je kot agregacijska funkcija v tem primeru uporabljena vsota (Sum).



Slika 4.9: Vmesnik SSDT za določanje strukture OLAP kocke



Slika 4.10: Vmesnik SSDT za urejanje dimenzije

V levem spodnjem razdelku slike 4.9 so navedene dimenzije. Za dimenzije določimo najprej vir podatkov. V primeru čiste zvezdne sheme podatkovnega skladišča je to kar posamezna dimenzijska tabela. Kako se ureja dimenzijo v vmesniku orodja SSDT, vidimo na sliki 4.10. V levem razdelku ('Attributes') izberemo atribute, katere želimo, da so vsebovani v dimenziji in po katerih bomo lahko analizirali podatke.

Sestavimo lahko tudi več hierarhij (osrednji razdelek 'Hierarchies'), ki so zelo priročen in uporabniku naraven način združevanja oz. razčlenjevanja podatkov. Primer na sliki 4.10 v sredini prikazuje hierarhijo, ki smo jo zgradili za dimenzijo 'Omrežje'. Ta hierarhija združuje zapise iz dimenzijske tabele 'D_Omrezje' po nivojih, opisanih v točki 4.2.6. V tej hierarhiji smo uporabili nivo 1 (RTP ali RP), nivo 3 (izvod), nivo 6 (TP), nivo 7 (razdelilec) in nivo 8 (izvod razdelilca ali celica).

	Measure Groups		
Dimensions	F Energija	F Energija MM	F Energija TP
Čas	Datum	Datum	Datum
Števci (Števec)	ID Stevec	ID Stevec	
Pogodba o dostopu (Pogodba o uporabi ...	ID Pogodba O Dostopu	ID Pogodba O Dostopu	
D Merilno Mesto (Merilno mesto)	SMM	SMM	
D TP (TP)			TP
Omrežje	ID Omrezje	ID Omrezje	ID Omrezje
D Kolicina (Količina)	ID Kolicina	ID Kolicina	ID Kolicina
D Dobavitelj (Dobavitelj)	ID Dobavitelj	ID Dobavitelj	
D Baza AMI (Baza AMI)	Db	Db	
D Da Ne (Je Obracun)		ID	
D Da Ne (Je Meritev)		ID	
D Veljavnost Pogodbe (Veljavnost Pogod...		ID Veljavnost Pogodbe	
D Cluster	ID Cluster	ID Cluster	

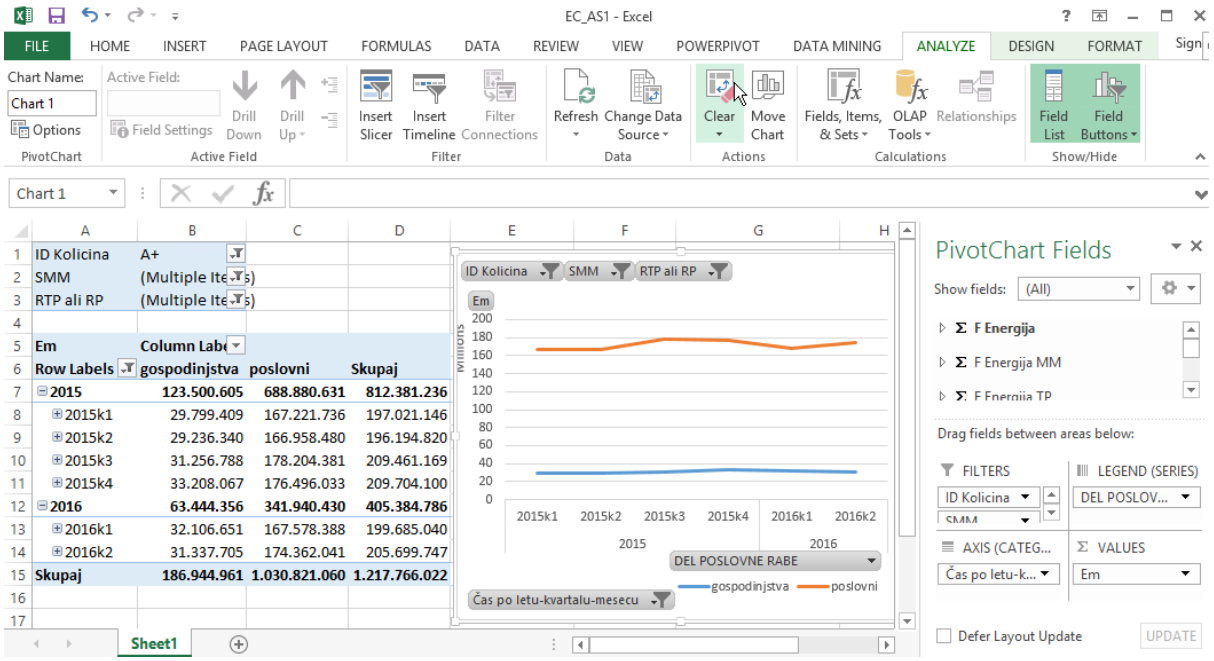
Slika 4.11: Vmesnik SSDT za določanje uporabe dimenzij v povezavi s skupinami mer

Ko si pripravimo vse dimenzije in mere, sledi določitev matrike uporabe dimenzij. Z uporabo SSDT in vmesnika, prikazanega na sliki 4.11, določimo katere dimenzije bodo uporabljene v povezavi s posamezno skupino mer in preko katerega atributa naj se izvede povezava.

Določanje teh povezav je zopet enostavno v primeru čiste zvezdne sheme podatkovnega skladišča. Takrat so povezave neposredne, preko ključev dimenzijskih tabel. Mehanizem sicer ponuja tudi kompleksnejše možnosti povezav, kot npr. referencirano povezavo (preko posredne dimenzijske tabele) ali povezavo mnogo-z-mnogo (dimenzijsko tabelo in tabelo dejstev povežemo preko druge tabele dejstev, ki je povezana s prvo tabelo dejstev preko druge dimenzijske tabele). Vendar moramo biti pri kompleksnejših vrstah povezav zelo previdni, saj se kaj hitro zgodi, da dobimo rezultate, ki so na prvi pogled povsem nelogični.

Po zaključeni pripravi modela kocke sprožimo procesiranje. V fazi razvoja in spreminjanja strukture to počnemo iz razvojnega orodja SSDT, kjer tudi spremljamo napredek in odkrivamo morebitne napake. Ko je procesiranje končano, je kocka objavljena na SSAS analitičnih storitvah strežnika SQL. Na ta podatkovni vir se lahko povežemo z raznimi OLAP odjemalci. Ena možnost za pregledovanje vsebine je kar v razvojnem orodju SSDT. Druga možnost za predstavitev podatkov uporabnikom je, da pripravimo poročila z bolj ali manj fiksno strukturo, ki pa jih je mogoče konfigurirati s parametri. Poročila objavimo na SSRS - poročevalskih storitvah strežnika SQL, do katerih uporabniki dostopajo s spletnim brskalnikom.

Pogosto uporabljen 'težki' odjemalec za OLAP podatke analitičnih storitev je Excel, aplikacija za preglednice iz zbirke Microsoft Office. Podatkovni vir povežemo preko vmesnika za analitične storitve strežnika SQL, ki ga izberemo iz nabora možnosti. V preglednici OLAP podatke pogosto predstavimo kot vrtilno (pivot) tabelo, ali jih vizualiziramo z grafom. Odjemalci kot je Excel, ponujajo uporabniku veliko možnosti in svobode pri izvajanju analiz.



Slika 4.12: Primer analize OLAP podatkov z odjemalcem Excel

Slika 4.12 prikazuje primer enostavne uporabe, ki v vrtilni tabeli in linijskem grafu podaja vsoto porabe delovne energije, izmerjene s sistemom AMI, po kvartalih leta 2015 in prvega polletja 2016. Čeprav je osnova za izračun sumarnih vrednosti več deset milijonov zapisov, dobimo rezultat v nekaj sekundah.

4.4 Odkrivanje značilnih dnevnih diagramov s podatkovnim rudarjenjem

Z izdelavo OLAP kocke smo s tem, ko smo v model analitične strukture vgradili pravila povezovanja oz. odnosov med obravnavanimi koncepti, dvignili nivo našega dojemanja od golih podatkov na raven informacij (glej sliko 2.2).

Naslednja raven razumevanja vsebin je znanje. O tem lahko govorimo takrat, ko poznamo in razumemo vzorce, ki se pojavljajo v podatkih. V našem primeru želimo raziskati dinamiko dnevne porabe odjemalcev električne energije oz. obliko dnevnega obremenilnega diagrama in v skupine združiti takšne s podobno dinamiko porabe. Zanima nas koliko je takšnih skupin diagramov, ki vsebujejo zadostno število elementov, da lahko skupino smatramo za tipično. Nadalje nas zanima, katere lastnosti odjemalcev so tiste, ki najbolj vplivajo na njihovo dinamiko porabe in na to, kateri odjemalci se bodo uvrstili v posamezno tipično skupino.

Te naloge se bomo lotili z metodami strojnega učenja in sicer s tistim segmentom le tega, ki mu popularno rečemo podatkovno rudarjenje (angl. data mining). Najprej bomo nad reprezentativnim vzorcem odjemalcev, za katere obstajajo podatki o 15-minutnih porabah, s postopki razvrščanja (angl. clustering) skušali odkriti, kakšne so tipične oblike diagramov dnevne porabe in na koliko skupin je diagrame smiselno razvrščati. To nalogo bomo opravili z različnimi algoritmi in različnimi vhodnimi parametri ter iz primerjave rezultatov ugotovili, kateri algoritem razvrščanja je najprimernejši za naše podatke ter kakšno je smiselno število skupin. Z algoritmom in vhodnimi parametri, ki se bodo izkazali za najboljše, bomo v tipične skupine razvrstili vse zapise o dnevnih porabah odjemalcev, za katere obstajajo podatki o 15-minutnih porabah. Tako bomo dobili skupine podobnih dnevnih diagramov porabe. Eno merilno mesto se bo s svojo porabo lahko en dan uvrstilo v eno skupino, drug dan pa v drugo. Z analizo števila razvrstitev diagramov za posamezno merilno mesto v različne skupine tipične oblike dnevne porabe bomo nato skušali merilnim mestom pripisati njihovo prevladujočo obliko porabe. Istočasno bomo skušali primerjati značilne oblike porabe med seboj in njihove morebitne spremembe skozi čas ter v tem odkriti kakšno pravilo.

Naslednji korak bo preizkus različnih metod nadzorovanega strojnega učenja, s katerimi bomo iskali korelacije med razvrstitvijo merilnega mesta v njegovo prevladujočo skupino ter različnimi parametri, kot so količina porabe, moč priključka, dejavnost odjemalca, določila pogodbe o dostopu in podobno. Z odločitvenimi drevesi (angl. decision trees) in povezovalnimi pravili (angl. association rules) bomo skušali ugotoviti, katere so tiste lastnosti, ki najbolj vplivajo na dnevno dinamiko porabe merilnega mesta oz. odjemalca, ki to merilno mesto uporablja in posledično na uvrstitev merilnega mesta v določeno skupino tipične dnevne porabe. Ugotavljali bomo katera metoda je najučinkovitejša pri uvrščanju v znane skupine. Z izbrano metodo bomo nato razvrstili vse preostale odjemalce, za katere ne obstajajo podatki o 15-minutnih porabah.

4.4.1 Izbira metode za razvrščanje daljinsko odbiranih merilnih mest v skupine tipične porabe

V poglavju 3 smo opisali nekaj teoretičnih osnov in dve metodi razvrščanja. Ti metodi bomo preizkusili z našimi podatki. Uporabili bomo Microsoftovi implementaciji algoritmov, ki sta že vgrajeni v SSAS:

- Metoda K-voditeljev (angl. K-means clustering) je najbolj klasična metoda razvrščanja z delitvenim postopkom, ki se v določenih pogojih (odvisno od vhodnih podatkov), lahko

zelo dobro odreže. Vendar ima algoritem nekaj slabosti, ki smo jih opisali v točki 3.4.1. Največja med njimi je občutljivost na posebneže tj. elemente množice, ki ležijo daleč od ostalih. Algoritem razvrsti vsak element množice vedno v eno samo skupino. V SSAS lahko izbiramo med dvema implementacijama, neskalabilno in skalabilno. Prva vedno upošteva celotno vhodno množico podatkov, medtem ko druga obravnava podmnožice s po 50.000 zapisi in obdelave ponavlja z dodajanjem podmnožic, dokler ne doseže želene stopnje konvergence. Običajno se to zgodi že v prvi ponovitvi.

- Metoda maksimiranja pričakovanj (angl. EM clustering - Expectation-Maximization), ki temelji na verjetnostnem pristopu k razvrščanju. Tudi ta implementacija enako kot K-means ponuja dve opciji izvajanja algoritma, neskalabilno in skalabilno. Skalabilna EM je hitrejša, še posebej pri velikih množicah podatkov, medtem ko v natančnosti ni praktično nič slabša od neskalabilne. Zato je tudi skalabilna EM v splošnem izbrana kot Microsoftova privzeta metoda za razvrščanje podatkov [5].

Pri obeh algoritmih kot vhodni podatek podamo število skupin, v katere naj se elementi razvrščajo. Ker vnaprej ne vemo kakšno je ustrezno 'naravno' število skupin, bomo razvrščanje ponovili za različna števila. Iz poznavanja narave odjema vemo, da se dinamika odjema oz. oblika dnevnega diagrama dokaj razlikuje med delovnimi in dela prostimi dnevi. To se dobro vidi tudi na slikah 2.4 in 2.5. Zato smo se odločili, da razvrščanje opravimo ločeno za delovne dni in ločeno za dela proste dni. Največje število skupin še sprejemljivih za bodočo analitsko uporabo smo ocenili na 12. Tako smo določili, da bomo s testnim vzorcem podatkov izvedli razvrščanje z uporabo štirih različnih modelov za podatkovno rudarjenje. Modele smo označili s številko in oznako metode: 1 - skalabilna EM, 2 - neskalabilna EM, 3 - skalabilna K-means in 4 - neskalabilna K-means. Razvrščanje smo ponovili 11 krat, za K = 2 do 12. Pri vsaki razvrstitvi bomo ocenili kompaktnost skupin tako, da bomo izračunali evklidsko razdaljo med vsakim posameznim vzorcem in srednjo vrednostjo skupine, kateri pripada. Vsoto evklidskih razdalj vseh elementov testne množice do centrov njihovih skupin smo privzeli kot mero za uspešnost razvrščanja. Manjša kot je ta vsota, bolj kompaktne so skupine.

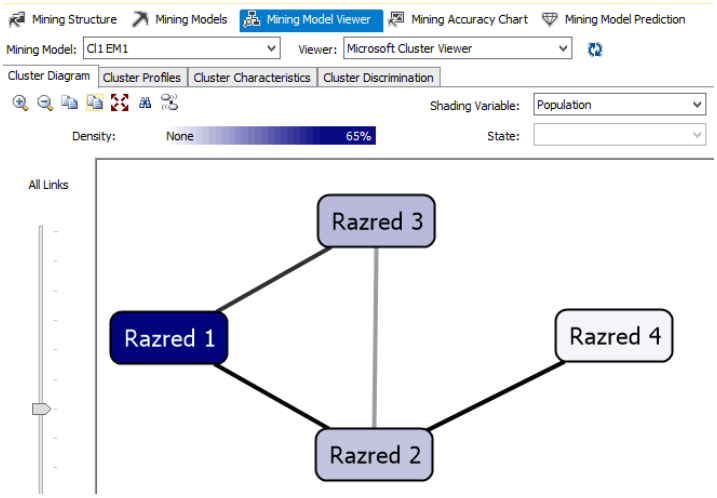
Za izvedbo primerjave smo si najprej pripravili testne podatke. Osnovni vir je bila tabela dejstev 'F_Energija' iz našega podatkovnega skladišča, opisana v točki 4.2.1.1. Posamezen zapis v tabeli združuje 96 vrednosti meritev ene vrste energije, normiranih na dnevno količino za en števec za en dan. Za nadaljnjo obdelavo smo iz tega vira izbrali le podatke o oddani delovni energiji (A+) za prvi kvartal leta 2016 in še to le s tistih števcev, za katere imamo podatke za vse dni obravnavanega obdobja in jim manjkata do največ dva 15-minutna odbirka v posameznem dnevu. S tem izločimo nepopolne podatke, katerih delež je sicer majhen, a bi kljub temu lahko precej pokvaril rezultat odkrivanja naravnih skupin in razvrščanja. Če bi uporabili vse zapise odjemalcev, ki ustrezajo tem kriterijem za vse dni obravnavanega obdobja, bi bilo teh zapisov okoli $5x10^6$, kar je pri takšnem številu ponovitev precejšen zalogaj za procesiranje modelov in nato izvedb razvrščanj. Zato smo se odločili izdelati manjši, a reprezentativen vzorec. Iz nabora vseh merilnih mest, ki so ustrezala navedenim pogojem, smo jih izbrali le približno 5%. Da je vzorec res reprezentativen smo zagotovili tako, da smo merilna mesta razporedili po skupni količini porabljene energije v obdobju in nato izbrali vsak dvajseti zapis.

Dobili smo 2.044 merilnih mest za katera je bilo v tabeli F_Energija za prvi kvartal 2016 na delovne dni za energijo A+ 128.769 zapisov. Naš vzorec je torej predstavljala množica 128.769 96-razsežnih elementov.

Izvedba razvrščanja poteka v dveh korakih. V prvem izvedemo izgradnjo modela s testno množico in z vhodnima parametroma 'metoda' in 'število skupin'. Ko je model zgrajen, lahko z uporabo tega modela in t.i. prediktivne poizvedbe neko množico podatkov ustrezne strukture razvrstimo v toliko skupin, kolikor jih je bilo predvidenih v modelu. V našem primeru smo vsakič razvrstili testno množico in si zapisali rezultate razvrščanja.

Te korake lahko prožimo ročno kar v razvojnem okolju SSDT. Najprej si na podlagi vira podatkov definiramo strukturo za rudarjenje. To storimo tako da povemo, kateri atribut predstavlja enovit identifikator zapisa in kateri izmed atributov naj bodo upoštevani pri rudarjenju kot spremenljivke. Spremenljivke so lahko zveznega ali diskretnega tipa. V naslednjem koraku določimo modele za rudarjenje, v našem primeru prej omenjene štiri, in nastavimo njihove parametre. Sedaj lahko poženemo procesiranje strukture za rudarjenje in modelov.

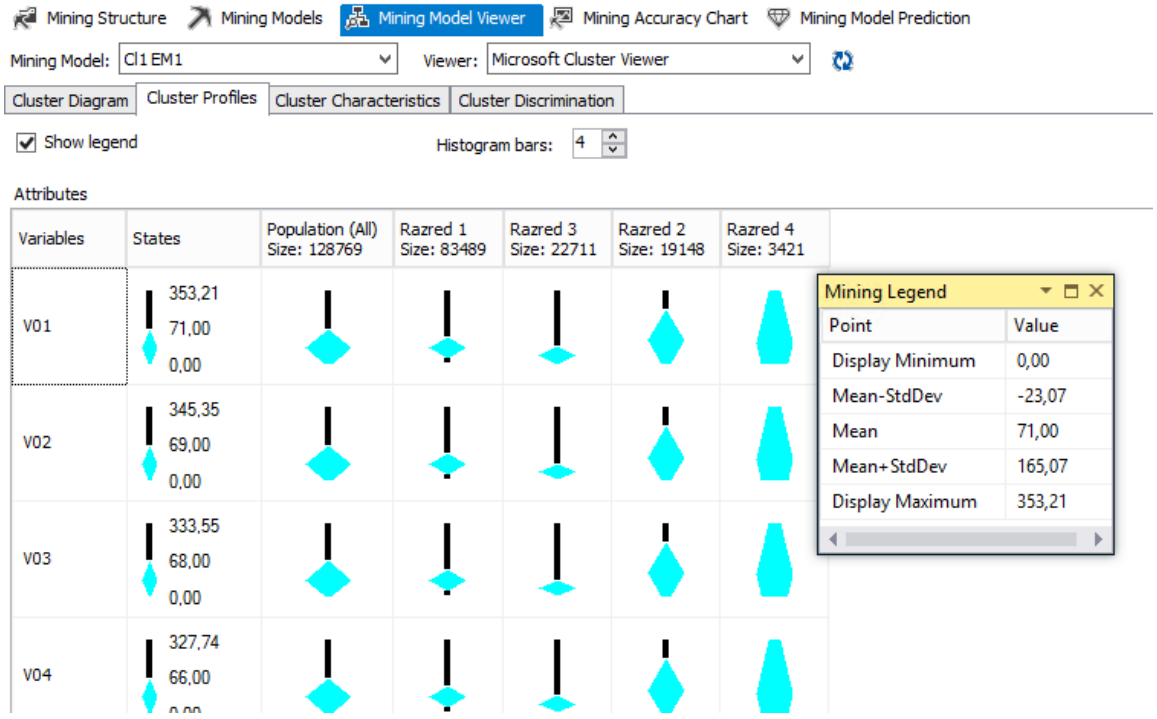
Rezultate si lahko ogledamo na več načinov. V nadaljevanju sta na slikah 4.13 in 4.14 prikazani obliki predstavitve rezultatov le za skalabilno metodo EM in za število skupin K=4.



Slika 4.13: Pregled lastnosti modela rudarjenja v diagramu skupin

V diagramu skupin na sliki 4.13 močneje obarvane skupine predstavljajo skupine z več člani. Podobnost med skupinami je ponazorjena z debelino povezav med njimi. Z drsnikom ob strani lahko spreminjamo nivo barvanja povezav, da vidimo le najmočnejše ali pa prikažemo tudi šibkejše povezave.

V profilih skupin na sliki 4.14 si lahko v grafični ponazoritvi ogledamo pravila za razvrščanje v posamično skupino glede na posamično spremenljivko.



Slika 4.14: Pregled lastnosti modela rudarjenja v obliki profilov skupin

Podobno informacijo dobimo v besedilni obliki v generičnem pregledovalniku, katerega del prikazuje slika 4.15.

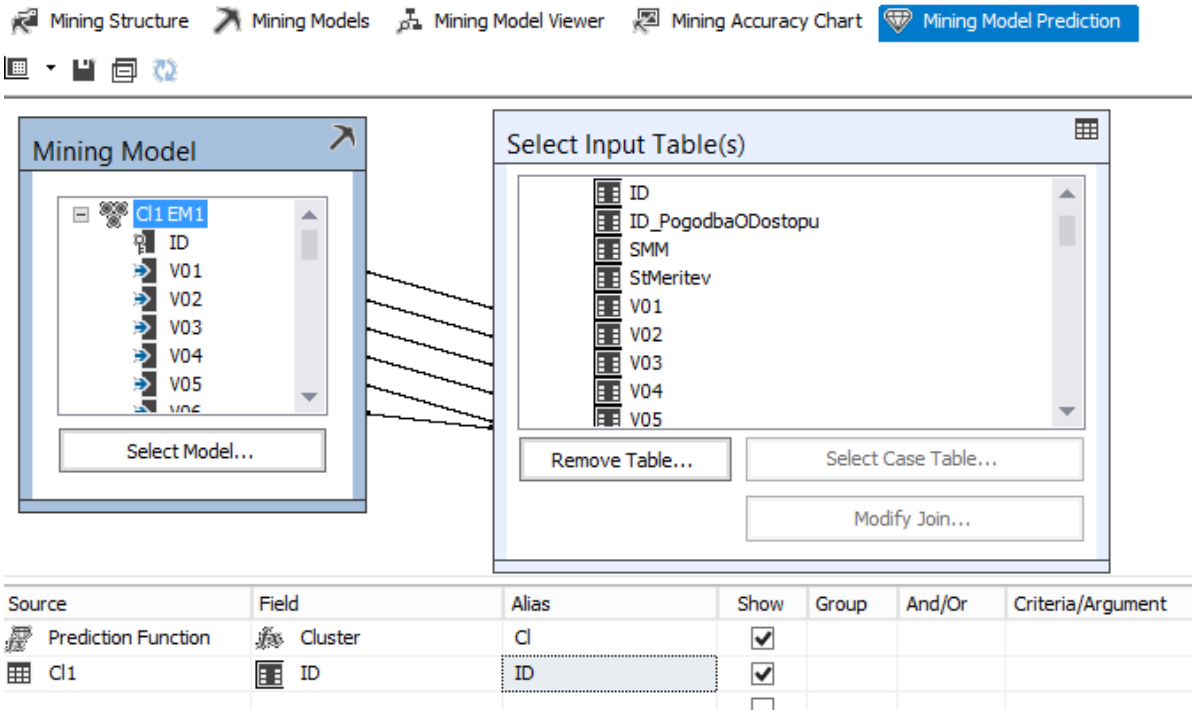
Node Details

MODEL_CATALOG	EC_A1
MODEL_SCHEMA	
MODEL_NAME	Cl1 EM1
ATTRIBUTE_NAME	
NODE_NAME	001
NODE_UNIQUE_NAME	001
NODE_TYPE	5 (Cluster)
NODE_GUID	
NODE_CAPTION	Razred 1
CHILDREN_CARDINALITY	0
PARENT_UNIQUE_NAME	000
NODE_DESCRIPTION	0 <=V19 <=195 , 0 <=V20 <=207 , 0 <=V21 <=219 , 0 <=V18 <=185 , 0 <=V84 <=275 , 0 <=V86 <=272 , 0 <=V22 <=228 , 0 <=V25 <=263 , 0 <=V17 <=174 , 0 <=V85 <=270 , 0 <=V26 <=274 , 0 <=V87 <=278 , 0 <=V23 <=244 , 0 <=V83 <=285 , 0 <=V82 <=291 , 0 <=V27 <=278 , 0 <=V88 <=269 , 0 <=V24 <=257 , 0 <=V16 <=167 , 0 <=V96 <=182 , 0 <=V89 <=262 , 0 <=V12 <=157 , 0 <=V13 <=158 , 0 <=V10 <=155 , 0 <=V80 <=313 , 0 <=V81 <=302 , 0 <=V95 <=190 , 0 <=V14 <=161 , 0 <=V15 <=162 , 0 <=V94 <=199 , 0 <=V68 <=325 , 0 <=V90 <=247 , 0 <=V09 <=155 , 0 <=V11 <=155 , 0 <=V01 <=175 , 0 <=V28 <=283 , 0 <=V07 <=157 , 0 <=V04 <=162 , 0 <=V79 <=324 , 0 <=V93 <=210
NODE_RULE	
MARGINAL_RULE	
NODE_PROBABILITY	0,806196237855167

Slika 4.15: Del generičnega pregledovalnika lastnosti modela podatkovnega rudarjenja

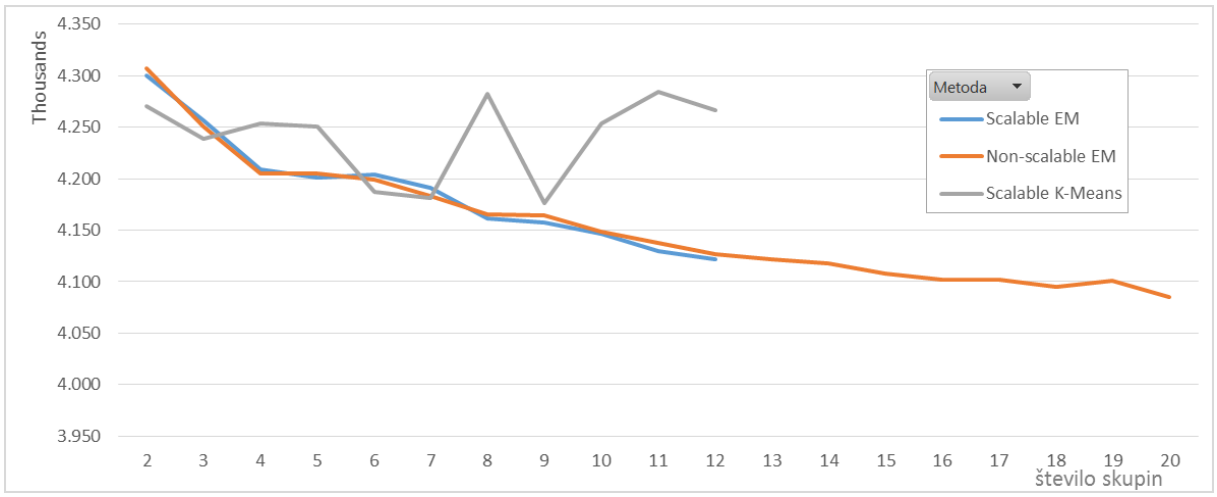
Iz grafičnega uporabniškega vmesnika razvojnega okolja SSDT lahko pripravimo in izvedemo tudi prediktivno poizvedbo, ki na podlagi izdelanega modela v skupine razvrsti podano vhodno množico elementov.

V nekaj ročnih ponovitvah procesiranja modelov z različnimi vhodnimi parametri in primerjavo rezultatov smo ugotovili, da sta metodi v modelih 3 (skalabilna K-means) in 4 (neskalabilna K-means) vedno vrnili povsem enake rezultate. Tudi njun čas procesiranja je bil enak. Zato smo metodo 4 izločili iz nadaljnje primerjave.



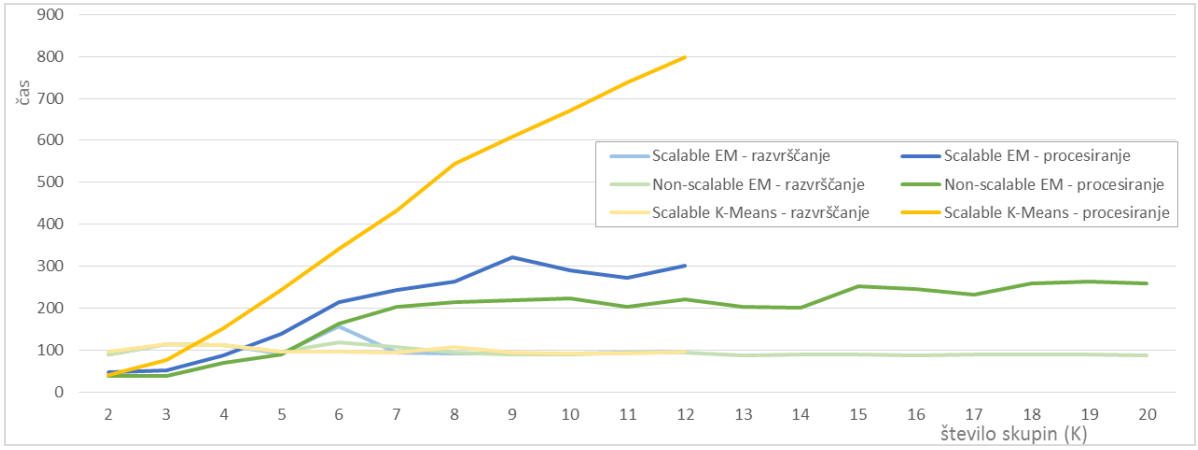
Slika 4.16: Urejanje in izvajanje prediktivne poizvedbe z orodjem SSDT

Ročno izvajanje procesiranja modelov podatkovnega rudarjenja za vse kombinacije različnih parametrov modelov kot smo si zamislili, bi bilo izjemno zamudno. Zato smo si postopek spreminjanja parametrov za modele, izvedbe procesiranja in razvrščanja testne množice podatkov avtomatizirali s pomočjo integracijskih storitev strežnika SQL (SSIS). V proceduro smo vgradili tudi merjenje časa procesiranja posameznega modela in časa izvedbe razvrščanja po posameznem modelu. Proceduro smo ponovili v zanki, za K=2, 3, … 12. Ob zaključku izvajanja smo vgradili še izračun evklidskih razdalj posameznih elementov do središča njihove skupine in vse te razdalje sešteli za posamezen model za ponovitev z določenim številom skupin. Tako smo dobili rezultate, ki z vsoto razdalj ponazarjajo razpršenost in kažejo na ustreznost razvrščanja. Na sliki 4.17 vidimo, da so rezultati za model z algoritmom K-means veliko slabši. Pri rezultatih tega algoritma smo tudi pogosto opazili, da je v posamezne skupine uvrstil zelo velik delež vzorcev, v druge pa izredno malo število vzorcev ali pa sploh ni naredil predpisanega števila skupin, ker je kakšna ostala prazna. Na sliki 4.19 vidimo koliko elementov je posamezen algoritem razvrstil v določeno skupino pri različnem skupnem številu skupin. Ta pojav, ko je npr. ena skupina zelo dominantna, v več preostalih skupinah pa je izredno malo članov, so povzročili 'osamelci' v obravnavani množici, na katere je algoritem K-means zelo občutljiv. Ta algoritem smo zato izločili iz nadaljnje obravnave. Za obe varianti algoritma EM so rezultati vsot razdalj zelo podobni. Obe krivulji razdalj imata izrazitejše koleno pri K=4. V tem delu grafa je nekoliko boljše rezultate dosegla neskalabilna metoda EM. Zanimivo pa je, da se rezultati kar nekoliko razlikujejo, ko primerjamo populacije skupin (zgornja dva grafa na sliki 4.19). Zato smo se odločili, da bomo v nadaljevanju uporabljali neskalabilno metodo EM. S testno množico podatkov in le z izbrano metodo smo nato izvedli še iteracije razvrščanj za K=13…20 in rezultate dodali v graf.



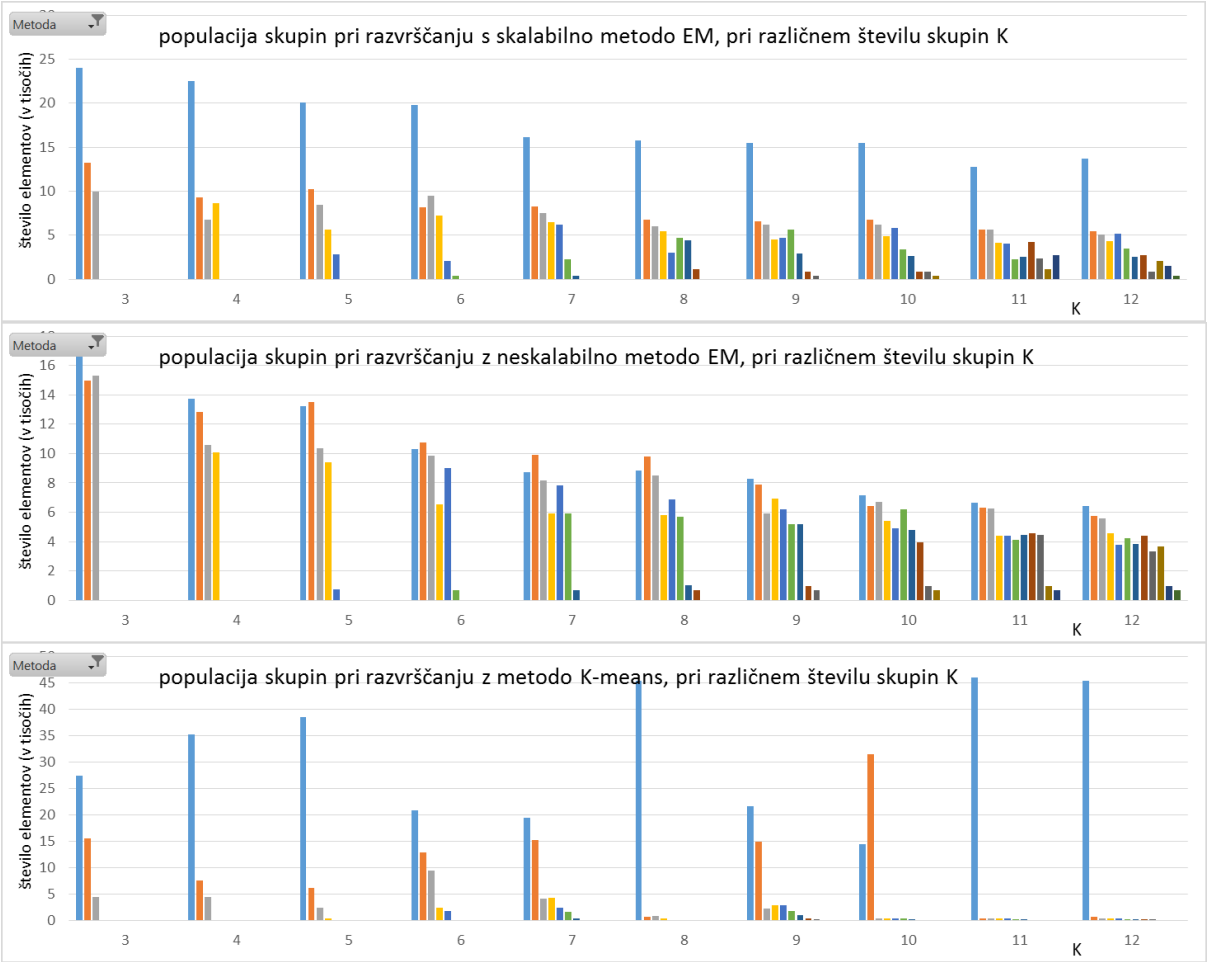
Slika 4.17: Vsote evklidskih razdalj testne množice v odvisnosti od števila skupin, za različne metode razvrščanja

Kot smo že omenili, smo ob izvajanju merili tudi čas procesiranja in čas razvrščanja. Tudi te rezultate smo zapisali, da lahko izvedemo primerjavo še po tej plati. Na sliki 4.18 vidimo, da je metoda K-means časovno potratnejša, s povečevanjem števila skupin pa se čas procesiranja modela skoraj linearno povečuje. Tudi zaradi tega je ta metoda manj primerna.



Slika 4.18: Čas procesiranja modela in čas razvrščanja za različne metode združevanja

Presenetilo nas je, da je bila neskalabilna metoda EM hitrejša od skalabilne. Tudi to dejstvo je pripomoglo, da smo dali prednost neskalabilni metodi EM. Pozitivno je tudi presenečenje, da z naraščanjem števila skupin od K=8 naprej čas potreben za procesiranje modela skorajda ni več naraščal.



Slika 4.19: Populacija skupin pri razvrščanju podatkov za delovnike prvega kvartala 2016 z različnimi modeli in za različno število skupin

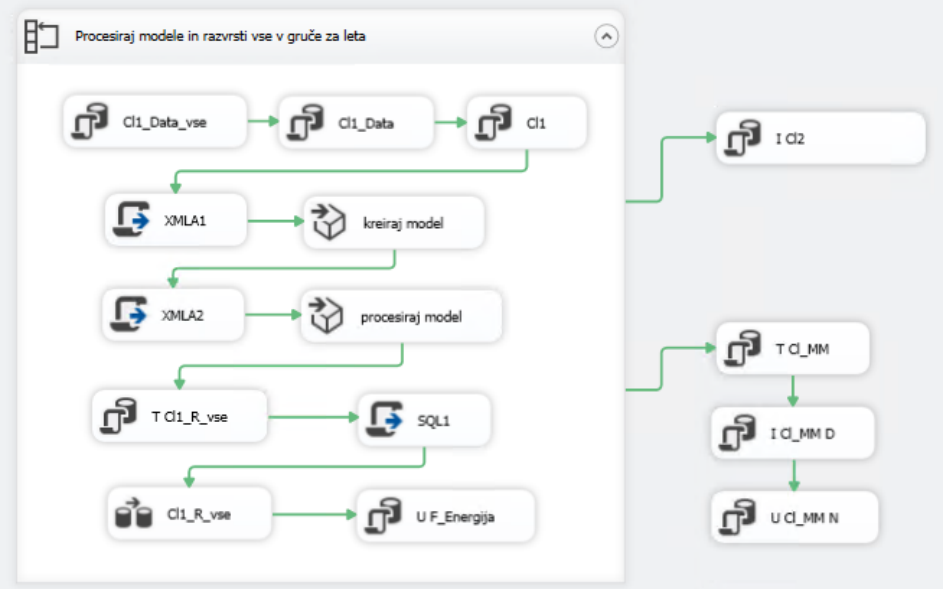
S primerjavo metod ob razvrščanju testne množice smo se torej odločili, da bomo v nadaljevanju za razvrščanje vseh dnevnih diagramov porabe merilnih mest s pametnimi števci uporabili neskalabilno metodo EM. Pri odločanju o številu skupin v katere bomo razvrščali, bi morda lahko razmišljali o večjem številu, saj se s povečevanjem le tega vsota razdalj dokaj konstantno manjša. Vendarle smo se odločili za K=4, kjer je tudi nekoliko izrazitejše koleno. Za manjše število skupin smo se odločili predvsem ob dejstvu, da nameravamo ločeno obravnavati delovne in dela proste dni.

4.4.2 Primerjava tipičnih dnevnih diagramov porabe

Naslednja naloga je bila razvrstiti vse zapise iz 'F_Energija' v skupine, ki združujejo podobne dnevne diagrame. En zapis predstavlja 96 točk porabe energije enega merilnega mesta v enem dnevu, normiranih na njegovo dnevno porabljeno količino, kar razumemo kot en dnevni diagram. Ker smo predvidevali, da se tipične oblike dnevne porabe spreminjajo glede na letni čas, smo razvrščanje opravili za vsak kvartal posebej. Omejili smo se na leto 2015 in prva dva kvartala leta 2016. Seveda smo si tudi tukaj pomagali z avtomatiziranim postopkom, izvedenim v obliki SSIS paketa. V tem postopku v zanki ponovimo serijo naslednjih aktivnosti:

- pripravimo reprezentativno podmnožico podatkov iz obdobja,
- kreiramo model podatkovnega rudarjenja za obdobje iteracije,
- izvedemo procesiranje modela s prej pripravljeno podmnožico,
- z izdelanim modelom razvrstimo celotno populacijo obdobja iteracije.

Postopek v zanki ponavljamo za vsak kvartal in to ločeno za delovnike in ločeno za dela proste dni. Tako za vsako celo leto dobimo 8 modelov, za vsak kvartal enega za delovnike in enega za dela proste dni.



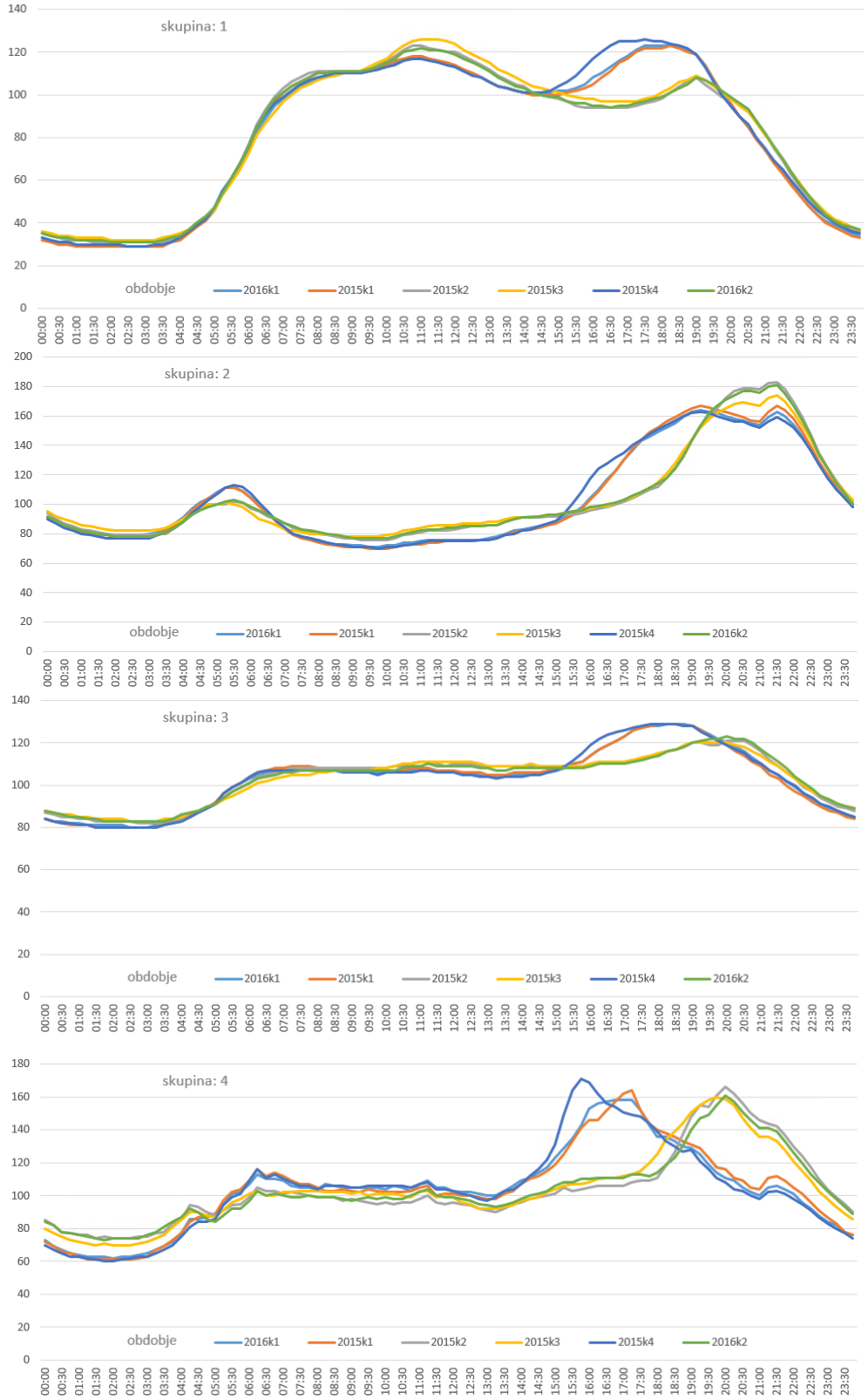
Slika 4.20: SSIS paket za postopek razvrščanja

Dobljene modele lahko primerjamo z medsebojno primerjavo parametrov (kot jih prikazujeta npr. slika 4.14 ali slika 4.15), vendar je takšna primerjava za človeško predstavo zelo težavna, še posebej, ker gre v našem primeru za 96-razsežni prostor.

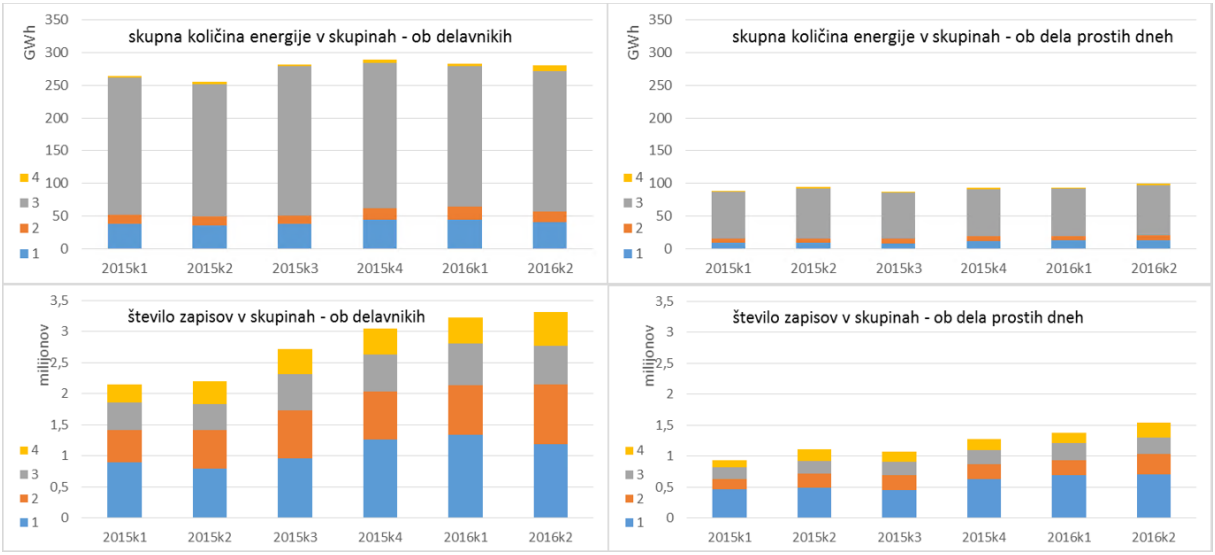
Zato smo raje naredili primerjavo rezultatov razvrščanja oz. primerjavo tipičnih dnevnih diagramov samih. Kot tipični dnevni diagram smo privzeli diagram, sestavljen iz srednjih vrednosti posameznih komponent vseh elementov, uvrščenih v isto skupino. Ob tem, da smo za vsako skupino izračunali srednje vrednosti za 96 točk normirane porabe in to zapisali kot posamezne zapise v tabelo tipičnih dnevnih diagramov (za vsak kvartal, delovne dni, dela proste dni), smo v zapis dodali še število elementov v skupini in povprečno dnevno količino energije.

Z analizo teh rezultatov lahko odkrijemo nekaj značilnosti tipičnih oblik porabe:

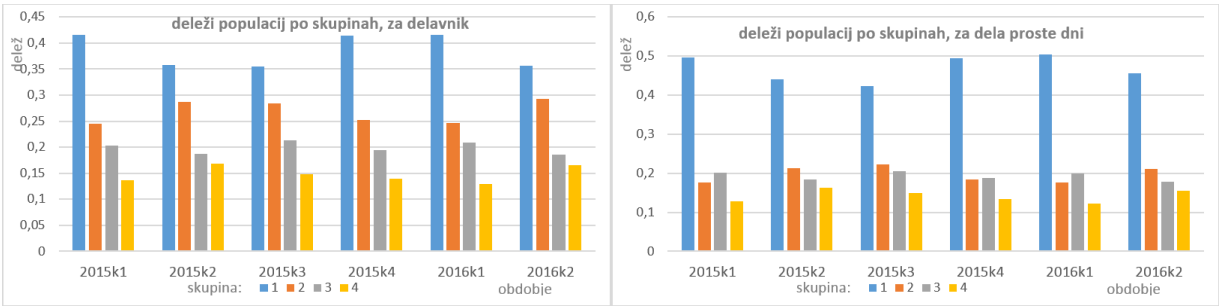
Pri diagramih vseh skupin opazimo, da se precej razlikujejo diagrami poletnih kvartalov (2015k2, 2015k3, 2016k2), od zimskih (2015k1, 2015k4, 2016k1) in sicer je na sliki 4.21 pri vseh opaziti popoldansko povečanje porabe pri zimskih kvartalih ob zgodnejših urah, kot pri poletnih kvartalih. Pri tem pa se oblika diagrama znotraj posamezne skupine skoraj popolnoma ujema za isti kvartal v različnih letih. Tudi deleži populacij v skupinah se od leta 2015 do 2016 niso bistveno spremenili kot je razvidno iz slike 4.23. Iz tega lahko zaključimo, da se navade odjemalcev iz leta 2015 do 2016 v povprečju niso spreminjale.



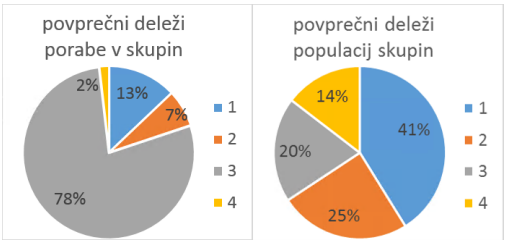
Slika 4.21: Tipični dnevni diagrami posamične skupine ob delovnikih, za kvartale



Slika 4.22: Skupne količine energije (zgoraj) in število zapisov (spodaj) v skupinah, ločeno za delovne (levo) in dela proste dni (desno), po kvartalih



Slika 4.23: Deleži populacij skupin po kvartalih, ločeno za delovnike in dela proste dni



Slika 4.24: Deleži količin porabe (levo) in števila elementov (desno) v skupinah, v povprečjih za vse kvartale in združeno za delovne in dela proste dni

Poleg navedenih splošnih ugotovitev ugotavljamo značilnosti skupin:

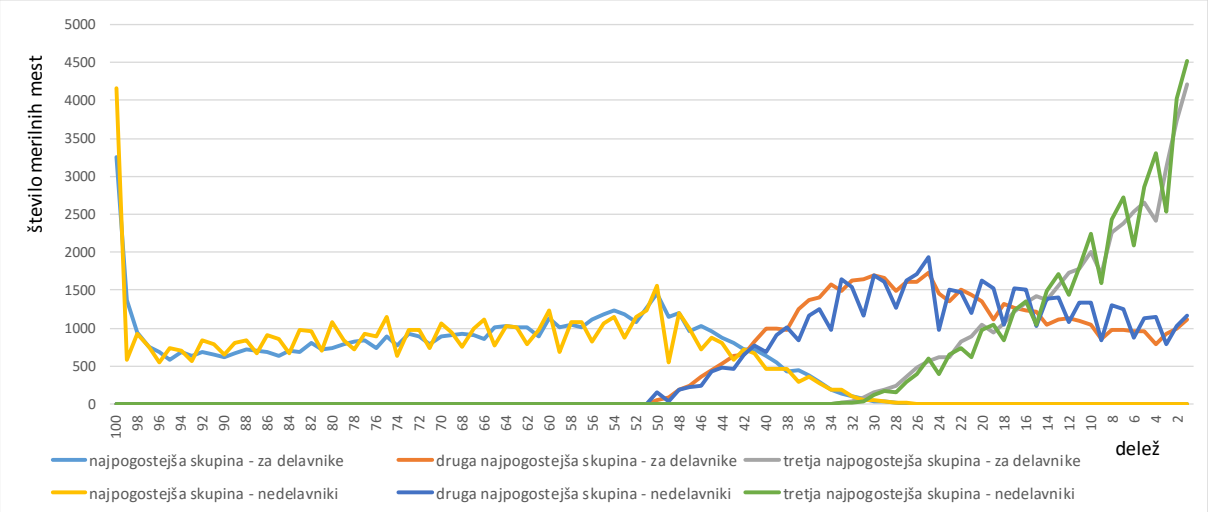
Skupina 1: Izrazito trapezna oblika (poraba ponoči je veliko manjša kot čez dan). V zimski sezoni je poraba pozno popoldne višja kot poleti. Povprečne količine porabe so nizke, saj 41% vseh elementov ustvari komaj 13% porabe.

Skupina 2: Cel dopoldan in ponoči je poraba nizka, zvečer izrazito naraste. Povprečna poraba je nizka, saj 25% populacije ustvari le 7% porabe.

Skupina 3: Najbolj konstantna poraba čez cel dan med vsemi oblikami. V skupini so v povprečju večji odjemalci, saj predstavljajo 20 % populacije, ki ustvari 78% porabe.

Skupina 4: Skupino sestavljajo v povprečju odjemalci z izrazito majhno porabo, saj 14% populacije ustvari 2% porabe. Večerna konica v diagramu porabe je bolj izrazita kot pri drugih skupinah. Med vrhovoma zimske in poletne večerne konice je kar 3-urni časovni zamik.

Pri razvrščanju vseh elementov smo vse zapise obravnavali enakovredno. Zapise istega merilnega mesta za različne dneve smo obravnavali kot medsebojno povsem neodvisne. Merilnih mest, katerih podatki so ustrezali postavljenim kriterijem celovitosti, je bilo nekaj več kot 59.000. Pričakovali bi, da se posamezen odjemalec drži dokaj ustaljenih navad pri koriščenju električne energije in da bo po dneh večinoma razvrščen v isto skupino. Ali to res drži in ali lahko za vsako merilno mesto enolično rečemo, kateri tipični obliki dnevnega diagrama oz. kateri skupini pripada, smo preverili tako, da smo za vsako merilno mesto pogledali v kateri skupini se pojavi največkrat, katera skupina je za to merilno mesto druga najpogostejša in katera tretja najpogostejša. Nato smo za vsako merilno mesto izračunali delež pojavljanja v njegovi najpogostejši skupini, delež pojavljanja v drugi najpogostejši in delež pojavljanja v tretji najpogostejši skupini. To smo naredili pri vsakem merilnem mestu ločeno za delovnike in ločeno za dela proste dni. V grafu na sliki 4.25 je prikazano število merilnih mest glede na delež pojavljanja v njihovi najpogostejši, drugi najpogostejši in tretji najpogostejši skupini.

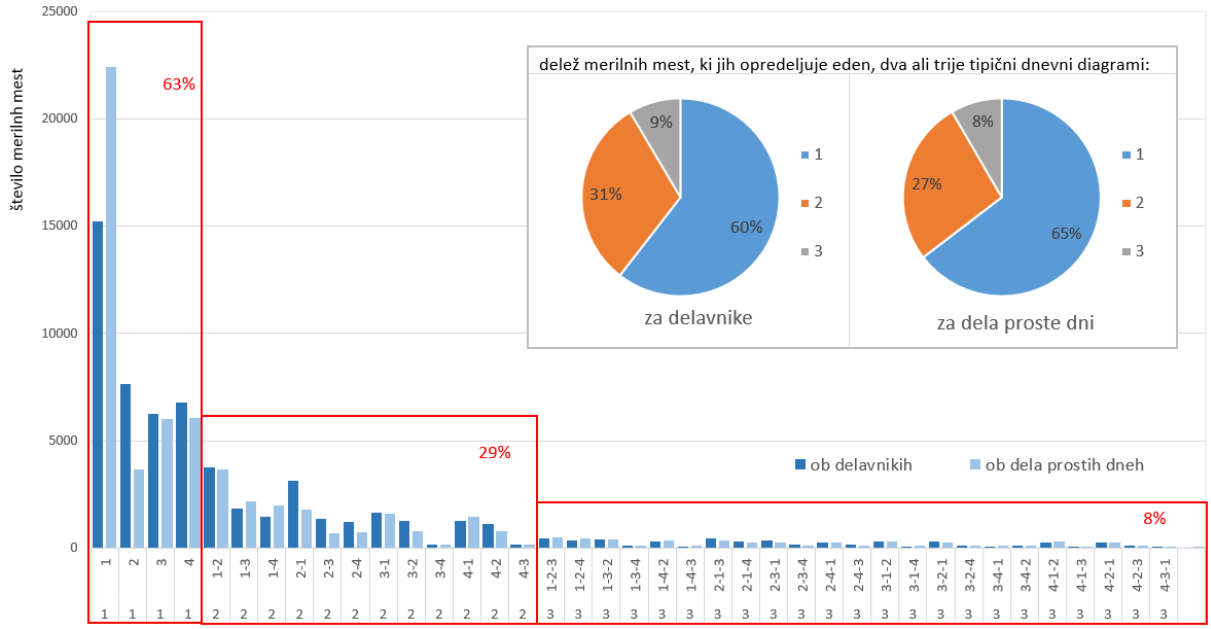


Slika 4.25: Število merilnih mest v odvisnosti od deležev pogostosti pojavljanja v svoji prvi, drugi in tretji najpogostejši skupini

Vidimo, da se razporeditve za delovne in dela proste dni držijo skoraj popolnoma enakih razmerij. Zato bomo pri določanju značilne skupine za merilno mesto uporabili enaka pravila za delovnike in za dela proste dni.

Iz razporeditvenih krivulj na sliki 4.25 smo se odločili, da bomo za merilna mesta, ki se v več kot 60 odstotkih vseh primerov uvrstijo v isto skupino rekli, da je to njihova prevladujoča skupina in da imajo enoznačno opredeljeno svojo značilno skupino. Za merilna mesta, ki

prvemu pogoju ne ustrezajo in se njihovi dnevni diagrami bolj razpršeno razvrščajo med dve ali več skupin smo se odločili, da bomo za tiste, ki se v več kot 75 odstotkih uvrstijo v svoji prvi dve najpogostejši skupini rekli, da jih opredeljujeta dve skupini, najprej tista z večjim deležem razvrstitev in nato tista z nižjim. Za vsa ostala merilna mesta bomo rekli, da jih opredeljujejo tri skupine in sicer v vrstnem redu pogostosti pojavljanja njihovih dnevnih diagramov v teh treh skupinah. Koliko merilnih mest se na tak način razvrsti v eno samo prevladujočo skupino, koliko med dve in koliko med tri skupine, vidimo na sliki 4.26.



Slika 4.26: Število merilnih mest, opredeljenih s posamezno skupino, s kombinacijo dveh ali s kombinacijo treh skupin tipičnih dnevnih porab

Kot vidimo, se z uporabljenim pravilom zelo velik delež merilnih mest s svojim dnevnim diagramom zelo enolično prepozna v enem samem tipičnem dnevnem diagramu.

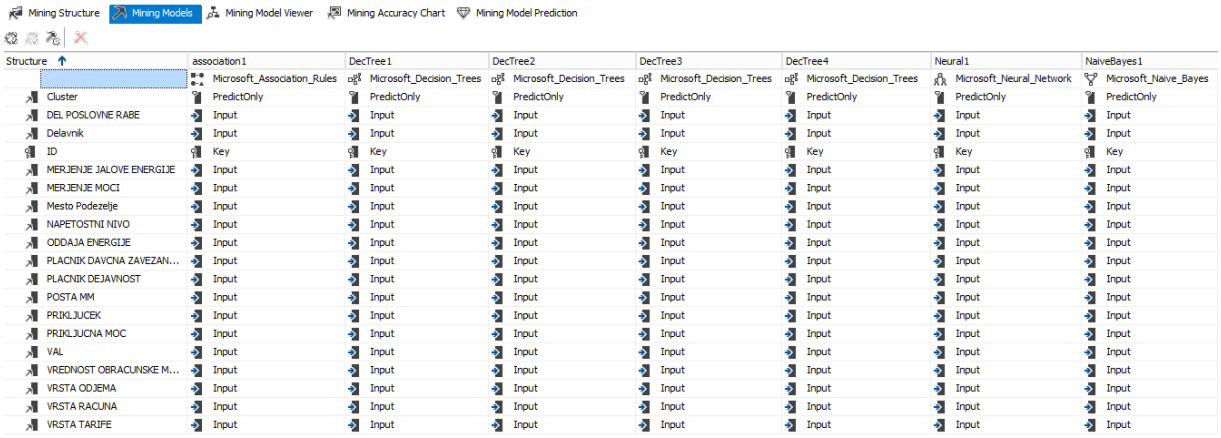
Če sedaj rečemo, da vsako merilno mesto razvrstimo enoznačno samo v tisto skupino, v katero je razvrščenih največ njegovih dnevnih diagramov, potem vemo, da vsaj za dve tretjini merilnih mest ta razvrstitev velja bolj ali manj ves čas. Le 8 odstotkov je takšnih merilnih mest prvotno okarakteriziranih s po tremi skupinami, ki dinamiko svoje porabe spreminjajo tako, da se razvrščajo zelo raznoliko in nestanovitno.

4.4.3 Uvrščanje merilnih mest, ki nimajo podatkov o 15-minutnih porabah, v skupine

V prejšnjih točkah smo opisali, kako smo z metodami razvrščanja (angl. clustering) odkrili, katere so značilne oblike dnevne porabe energije. V skupine smo po podobnosti dinamike dnevne porabe razvrstili vsa tista merilna mesta, za katera s sistemom naprednega merjenja (AMI) zbiramo podatke o 15-minutnih porabah in so bili podatki v obdobju od 1.1.2015 do 30.6.2016 celoviti in konsistentni. Takšnih merilnih mest je približno 57.000. V razvrščanje

nismo vključili približno 42.000 merilnih mest, ki so sicer vključena v sistem naprednega merjenja, a zanje zbiramo le urne meritve. V prejšnji točki smo analizirali značilne dnevne diagrame, sedaj pa nas zanima ali lahko pri merilnih mestih, ki so po obliki dnevne porabe razvrščena v isto skupino, najdemo kakšno korelacijo oziroma povezanost med lastnostmi, ki izhajajo iz pogodb o dostopu za odjemalce na teh merilnih mestih, njihove lokacije in časa. Tudi za izvedbo te naloge smo uporabili metode podatkovnega rudarjenja in sicer tokrat metode za klasifikacijo in predikcijo. Analitične storitve strežnika SQL (SSAS - SQL Server Analysis Services) nam ponujajo več algoritmov, s katerimi se lahko lotimo zadane naloge. Odločili smo se, da bomo preizkusili štiri različne algoritme, skušali najti optimalne nastavitve vhodnih parametrov za vsakega in nato s primerjanjem rezultatov med njimi izbrali najučinkovitejšega. Izbrali smo algoritme, ki temeljijo na metodi odločitvenih dreves (angl. decision trees), opisanih v točki 3.6.1, metodi povezovalnih pravil (angl. association rules), opisanih v točki 3.6.2, metodi nevronskih mrež in metodi z naivnim Bayesovim klasifikatorjem (naive-Bayes). S temi algoritmi smo izdelali več modelov za podatkovno rudarjenje.

Strukturo nad katero bomo izvajali podatkovno rudarjenje smo pripravili tako, da smo za vsa merilna mesta, ki smo jih v prejšnji točki razvrstili v skupine, ob podatku o razvrstitvi (atribut imenovan 'Cluster' na sliki 4.27), dodali tiste atribute, za katere smo menili, da lahko kakorkoli vplivajo na dinamiko porabe energije. Za posamezno merilno mesto smo podali po en zapis za vsak dan, s podatkom o dnevni količini porabljene energije in s podatkom o tem ali gre za delovnik ali dela prost dan. Za omenjeno obdobje leta in pol smo tako ustvarili množico z okoli 24 milijoni elementov. Podatek o razvrstitvi predstavlja spremenljivko, na podlagi katere bomo najprej zgradili model, nato pa preverili njegovo uspešnost. Strukturo s parametri (na levi strani) in modele za podatkovno rudarjenje, izdelane v razvojnem okolju SSDT, vidimo na sliki 4.27.


Mining Structure | Mining Models | Mining Model Viewer | Mining Accuracy Chart | Mining Model Prediction

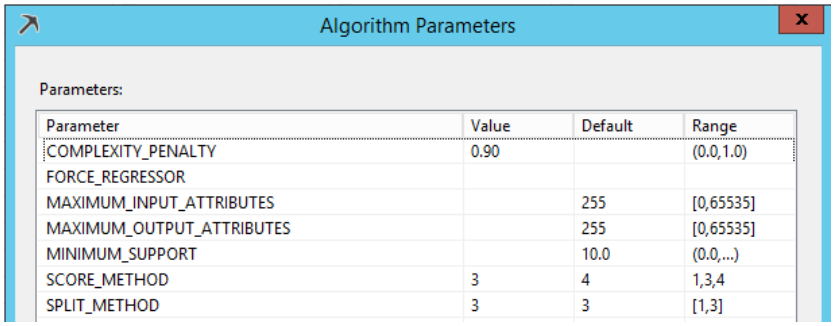
Structure	association1	DecTree1	DecTree2	DecTree3	DecTree4	Neural1	NaiveBayes1
	Microsoft_Association_Rules	Microsoft_Decision_Trees	Microsoft_Decision_Trees	Microsoft_Decision_Trees	Microsoft_Decision_Trees	Microsoft_Neural_Network	Microsoft_Naive_Bayes
Cluster	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly
DEL POSLOVNE RABE	Input	Input	Input	Input	Input	Input	Input
Delavnik	Input	Input	Input	Input	Input	Input	Input
ID	Key	Key	Key	Key	Key	Key	Key
MERJENJE JALOVE ENERGIJE	Input	Input	Input	Input	Input	Input	Input
MERJENJE MOCI	Input	Input	Input	Input	Input	Input	Input
Mesto Podezelje	Input	Input	Input	Input	Input	Input	Input
NAPETOSTNI NIVO	Input	Input	Input	Input	Input	Input	Input
ODDAJA ENERGIJE	Input	Input	Input	Input	Input	Input	Input
PLACNIK DAVCNA ZAVEZAN...	Input	Input	Input	Input	Input	Input	Input
PLACNIK DEJAVNOST	Input	Input	Input	Input	Input	Input	Input
POSTA MM	Input	Input	Input	Input	Input	Input	Input
PRIKLJUCEK	Input	Input	Input	Input	Input	Input	Input
PRIKLJUCNA MOC	Input	Input	Input	Input	Input	Input	Input
VAL	Input	Input	Input	Input	Input	Input	Input
VREDNOST OBRACUNSKE M...	Input	Input	Input	Input	Input	Input	Input
VRSTA ODJEMA	Input	Input	Input	Input	Input	Input	Input
VRSTA RACUNA	Input	Input	Input	Input	Input	Input	Input
VRSTA TARIFE	Input	Input	Input	Input	Input	Input	Input

Slika 4.27: Struktura in modeli za podatkovno rudarjenje, izdelani v razvojnem okolju SSDT

Microsoftova implementacija algoritma odločitvenih dreves omogoča izbiro različnih metod za izgradnjo drevesa. Algoritem ob vsaki iteraciji členitve drevesa določa najbolj relevantne atribute (angl. feature selection) tj. tiste z največjo informacijsko vrednostjo. Kot metodo določanja pomembnosti atributov izberemo eno izmed razpoložljivih metod: metodo Shannonove entropije, Bayesovo metodo s K2-predniki ali Bayes - Dirichletov ekvivalent z

uniformnimi predniki. Način kako naj algoritem razčleni podveje drevesa, lahko nastavimo na: binarno (ne glede na obseg vrednosti atributa po katerem se odločamo, algoritem vedno izdela dve podveji), popolno (veji se v do toliko podvej, kolikor je možnih stanj atributa) in avtomatsko (algoritem se v vsaki iteraciji sam odloča ali bo izbral binarno ali popolno vejanje).

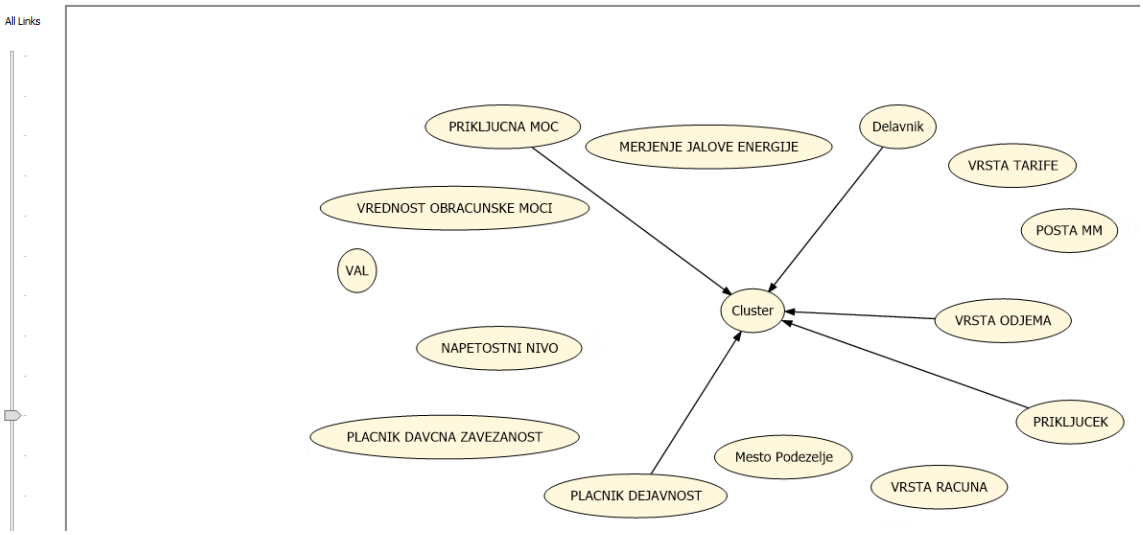
Vejanje odločitvenega drevesa je tako odvisno od izbrane metode za izbiro relevantnih atributov (parameter 'score_method'), načina vejanja (parameter 'Split_method'), tipov vhodnih spremenljivk in tipa izhodne spremenljivke.



Slika 4.28: Vmesnik SSDT za nastavitev parametrov za izdelavo modela odločitvenega drevesa

Za naš preizkus smo pripravili 7 različnih modelov (slika 4.27), od tega 4 z algoritmom odločitvenih dreves, z izbranimi različnimi metodami vrednotenja atributov in načina cepitve.

Po izvedeni izgradnji in procesiranju modelov podatkovnega rudarjenja lahko v grafičnem vmesniku orodja SSDT analiziramo izdelane modele. Načini podajanja (vizualizacije) modela so odvisni od izbranega algoritma rudarjenja. Za modele, izdelane po katerikoli metodi, razen z nevronskimi mrežami, lahko prikažemo t.i. mrežo odvisnosti, kot jo vidimo na sliki 4.29.



Slika 4.29: Pregled vpliva posameznih atributov na določitev ciljne spremenljivke

Z drsnikom ob levi strani diagrama spreminjamo 'občutljivost' za prikazovanje povezav, tako da zlahka ugotovimo, katere vhodne spremenljivke bolj vplivajo na izid odločitvenega drevesa

in katere manj. Za našo vhodno množico podatkov je metoda povezovalnih pravil določila naslednje zaporedje pomembnosti atributov: vrsta odjema, količina dnevno porabljene energije (spremenljivka 'VAL'), merjenje jalove energije, dejavnost plačnika, …

Metoda z naivnim Bayesovim klasifikatorjem nam da drugačno zaporedje pomembnosti atributov: priključek, priključna moč, dejavnost plačnika, vrsta odjema, …

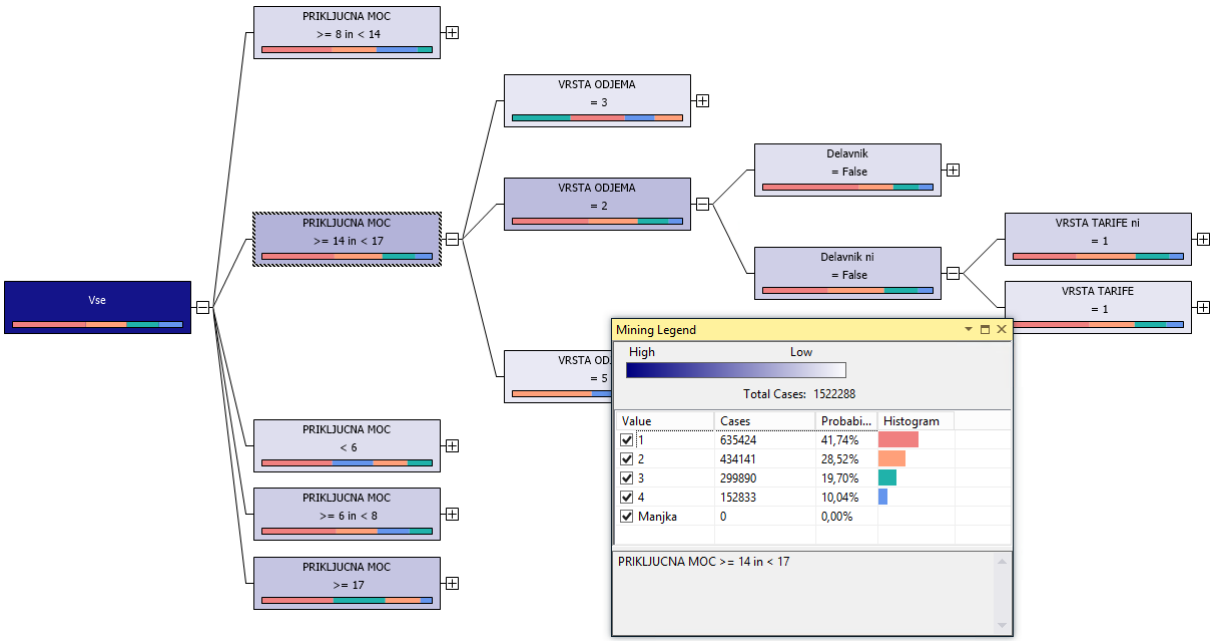
Pri metodah z odločitvenimi drevesi smo preizkusili varianto 'DecTree1' (Bayes s K2 predniki, polna delitev), ki je obtežila atribute v naslednjem vrstnem redu pomembnosti: priključna moč, vrsta odjema, dejavnost, priključek, delovnik, …

Enako začetno zaporedje atributov je izdelala varianta 'DecTree2' (Bayes-Dirichlet, polna delitev): priključna moč, vrsta odjema, dejavnost, priključek, delovnik, …

Precej drugačne uteži je pripravila varianta 'DecTree3' (Bayes-Dirichlet, binarna delitev) in sicer: vrsta odjema, priključna moč, obračunska moč, 'VAL', dejavnost plačnika, …

Zopet drugačno zaporedje atributov je izdelal algoritem v varianti 'DecTree4' (entropija, polna delitev), ki se je začelo s: priključna moč, dejavnost, pošta, delovnik, vrsta odjema, mesto-podeželje,…

Odločitveno drevo (angl. decision tree) zgrajeno s katerokoli metodo ocenjevanja atributov lahko nazorno pregledujemo v obliki drevesa, v katerem lahko razširjamo ali združujemo veje (slika 4.30). S takšnim prikazom je človeku zelo enostavno razumeti postopek odločanja in tudi obrazložiti rezultate uvrščanja. Pregledovalnik ponuja tudi prikaz statistike učne množice za izbrano vozlišče s številom primerov in njihovo verjetnostjo za vsako vrednost atributa vozlišča.

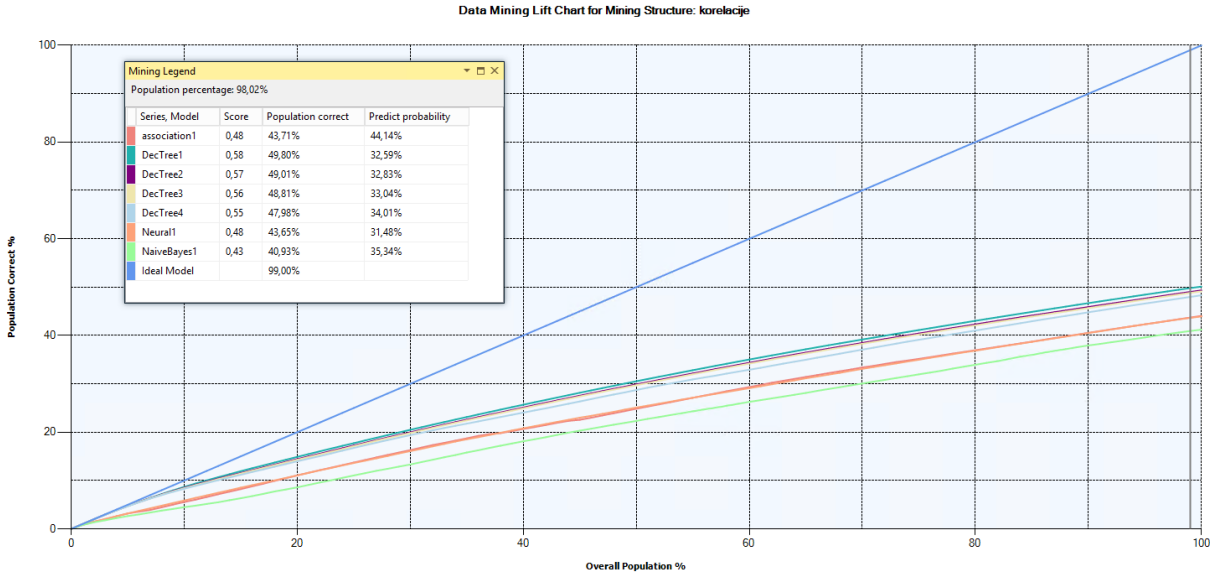


Slika 4.30: Pregledovalnik odločitvenega drevesa v okolju SSDT

Z odzivnim diagramom lahko tudi za odločitveno drevo, izdelano na podlagi naše vhodne množice ugotovimo, da ima v povezavi s tem, v kateri skupini je posamezno merilno mesto, največjo 'težo' priključna moč, vrsta odjema, dejavnost, priključek, delovnik, …

Na ravni izvedbe celotne obdelave vseh naštetih modelov podatkovnega rudarjenja smo nastavili, da 70% celotne množice predstavlja učno množico (angl. training set) in 30% testno množico ('HoldoutMaxPercent'=30).

S podatki učne množice mehanizem izdela modele za klasifikacijo, s podatki testne množice pa preveri njihovo pravilnost. Natančnost napovedi naših sedmih modelov nad testno množico lahko primerjamo na odzivnem diagramu (angl. lift chart). Ta diagram na abscisni osi prikazuje odstotek celotne populacije, na ordinatni osi pa odstotek pravilno napovedanih oz. uvrščenih elementov. Medsebojno primerjanje več modelov na tak način je možno, če napovedujejo isto spremenljivko.



Slika 4.31: Odzivni diagram za izdelane modele

Kadar, kot v našem primeru, ne določimo stanja napovedovane spremenljivke, nam diagram prikazuje obnašanje modelov za vsa stanja napovedovane spremenljivke. V tem primeru bi idealni model, ki vedno da pravilen odgovor, ponazarjala premica, ki povezuje točki (0%, 0%) in (100%, 100%). Med našimi modeli je torej najboljši tisti, čigar krivulja je najbližja premici idealnega modela.

Pri ponovitvah z različnimi množicami opazimo, da so modeli zelo občutljivi na spremembo vhodnih podatkov. Razmerja v uspešnosti med modeli se kaj hitro spremenijo. Da bi dobili stabilnejšo oceno uspešnosti algoritmov, smo izvedli križno preverjanje na podmnožicah (angl. K-fold cross validation). To je pogost način uporabljen v analitiki bodisi za preverjanje robustnosti posameznega modela bodisi za medsebojno primerjanje modelov in določitev najboljšega na podlagi statistike. Izvedemo ga tako, da učno množico razdelimo v K podmnožic in za vsako ponovimo postopek:

1. izbrano množico uporabimo za indukcijo pravil,
2. preostale podmnožice uporabimo za testiranje pravil.

Nad rezultati drugega koraka metoda izračuna statistične metrike: število pravilno klasificiranih primerov, število napačno klasificiranih primerov, 'Lift' – povprečje verjetnosti vseh napovedi za model, 'RMSE' - kvadratni koren povprečne kvadratne napake, deljen s številom primerov v podmnožici, povprečje logaritmov verjetnosti napovedi, ipd. Rezultati za 6-kratno križno preverjanje, razvrščeni padajoče po povprečnem številu pravilno klasificiranih primerov, so zbrani v tabeli 1.

	klasifikacija				verjetnost					
	nepravilno		pravilno		Lift		Log Score		RMSE	
model	povprečje	StD	povprečje	StD	povprečje	StD	povprečje	StD	povprečje	StD
DecTree1	19.020,4999	61,2175	14.312,8335	61,2819	0,1027	0,0015	-1,2350	0,0015	0,5769	0,0007
DecTree3	19.080,6672	43,9610	14.252,6662	43,5764	0,0993	0,0017	-1,2384	0,0017	0,5798	0,0010
DecTree2	19.093,1669	34,0708	14.240,1664	33,8500	0,0987	0,0019	-1,2390	0,0019	0,5800	0,0004
DecTree4	19.458,3335	46,5322	13.874,9998	46,4006	0,0838	0,0018	-1,2539	0,0018	0,5940	0,0003
NaiveBayes1	20.025,6670	69,7251	13.307,6664	69,5767	-0,2083	0,0061	-1,5460	0,0061	0,5013	0,0004
association1	20.198,8336	38,9120	13.134,4997	38,6815	0,1442	0,0009	-1,1935	0,0009	0,5605	0,0004
Neural1	20.296,8344	122,5806	13.036,4990	122,3048	0,0551	0,0037	-1,2826	0,0037	0,6072	0,0018

Tabela 1: Rezultati križnega primerjanja

Pravilnosti napovedanih razvrstitev so dokaj nizke, kar ne preseneča. Že vhodni podatki o razvrstitvi merilnih mest s 15-minutnimi porabami iz učne in testne množice, ki so bili v skupine razvrščeni z metodo EM za razvrščanje (točka 4.4.2), ne pripadajo skupinam povsem enoznačno. To je povsem naravno, saj bi težko trdili, da bi se morala dinamika dnevne porabe vsakega merilnega mesta natančno uvrščati v eno od štirih značilnih oblik. Poleg tega smo videli (točka 4.4.2, slika 4.26), da obstaja določen delež merilnih mest, ki spreminjajo naravo svoje porabe tako, da se enkrat uvrščajo v eno in drugič v drugo skupino. Glavni razlog, da rezultat ni boljši, pa je zagotovo v tem, da v resnici na dinamiko porabe električne energije poleg parametrov pogodbe o dostopu in dnevne višine porabe, vpliva še veliko drugih dejavnikov (kot npr. temperatura, osvetljenost, način ogrevanja in hlajenja, …), o katerih pa žal nismo imeli na voljo podatkov.

Po primerjavi rezultatov križnega primerjanja, smo izbrali model 'DecTree1', odločitveno drevo z Bayesovo metodo s K2-predniki ter polno delitvijo, ki se je najbolje odrezal v križnem preverjanju. S tem modelom smo izvedli prediktivno poizvedbo, s katero smo vsa merilna mesta, za katera nimamo podatkov o 15-minutnih porabah, za vsak dan celotnega obdobja, ki smo ga obravnavali, razvrstili v skupino z značilnim diagramom dnevne porabe, glede na lastnosti pogodbe o dostopu za posamezno merilno mesto ter glede na njegovo ocenjeno dnevno porabo, kot je opisano v točki 4.2.1.3.

4.4.4 Avtomatizirano izvajanje podatkovnega rudarjenja in procesiranja OLAP kocke

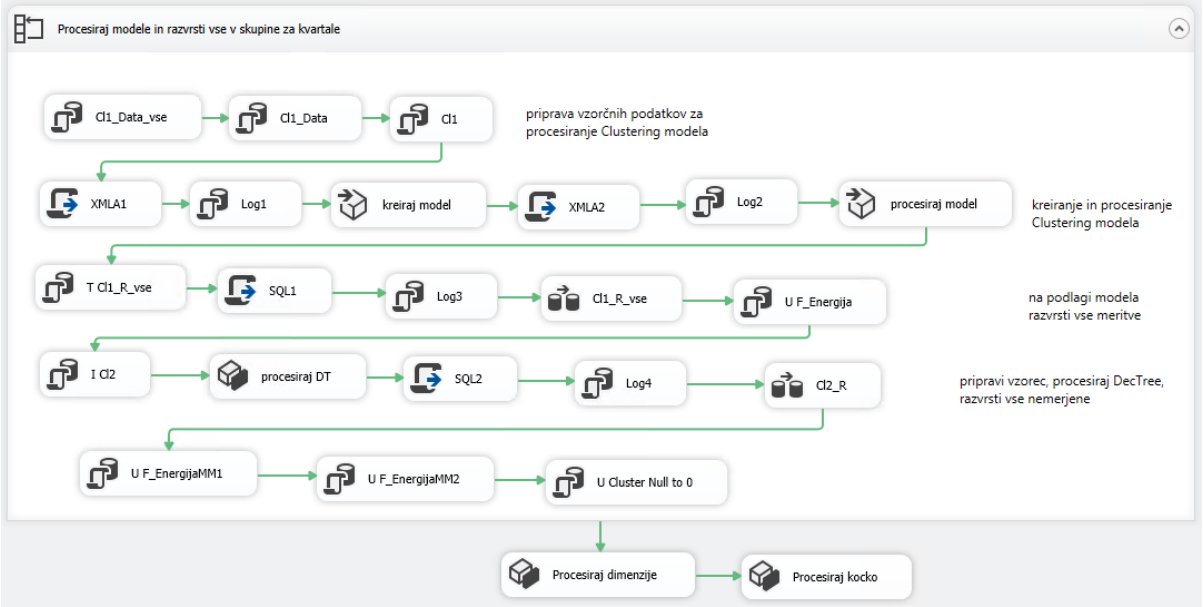
Podatek o razvrstitvi merilnega mesta v skupino z značilno obliko porabe je za nadaljnjo obravnavo pomemben, saj bomo na podlagi tega podatka predvidevali dnevno obliko porabe za merilna mesta, za katera nimamo dnevnih meritev. S tipičnim dnevnim diagramom (ločeno za

vsak kvartal in ločeno za delovnike in dela proste dni) in ocenjeno dnevno porabo posameznega merilnega mesta bomo zanj izdelali diagram predvidene dnevne porabe energije. To bomo naredili za celotno obdobje, zajeto v podatkovnem skladišču, za vsa merilna mesta, ki nimajo podatkov o 15-minutnih porabah. Torej bo potrebno v zanki za posamezne kvartale in ločeno za delovnike in dela proste dni ponavljati postopek:

1. Izdelaj in procesiraj model za razvrščanje v skupine (z neskalabilno metodo EM).
2. Na podlagi modela iz 1. koraka v skupine razvrsti vse dnevne diagrame za merilna mesta.
3. Z razvrstitvami dnevno merjenih merilnih mest in lastnostmi njihovih pogodb ter porabljenih količin energije z metodo odločitvenih dreves izdelaj model za klasifikacijo.
4. Z modelom iz 3. koraka vsem merilnim mestom brez dnevnih meritev za vsak dan (glede na pogodbo in ocenjeno dnevno porabo) pripiši zanje predviden tipični dnevni diagram porabe.

Ti podatki so podlaga za našo OLAP kocko, opisano v točki 4.3. Zaradi tega je potrebno po zaključenem vpisu podatkov izvesti še procesiranje OLAP kocke.

Celoten postopek smo ponovno implementirali kot integracijski paket SSIS. Prikazan je na sliki 4.32.



Slika 4.32: Grafična ponazoritev celotnega postopka razvrščanja merilnih mest in njihove dnevne porabe

Da bomo lahko OLAP kocko uporabljali za analize in poročanje pri rednem delu, bo potrebno vsebino redno posodabljati. Vsebine, ki smo jih uvrstili v podatkovno skladišče in v kocko so takšne narave, da bo zadoščalo osveževanje enkrat dnevno. Torej moramo po tem, ko s postopkom ETL dodamo v podatkovno skladišče dnevne spremembe (izvedemo SSIS paket, prikazan na sliki 4.8), izvesti še postopek razvrščanja s podatkovnim rudarjenjem (izvedemo SSIS paket, prikazan na sliki 4.32) za tekoči kvartal. Zadoščalo bi tudi, če bi model za razvrščanje v skupine (angl. clustering) in model za klasifikacijo (odločitveno drevo) izdelali

le enkrat tedensko in ju nato v dnevni obdelavi le uporabili za predikcijo. S tem bi znatno skrajšali čas izvajanja dnevne obdelave.

4.5 Primeri poročil

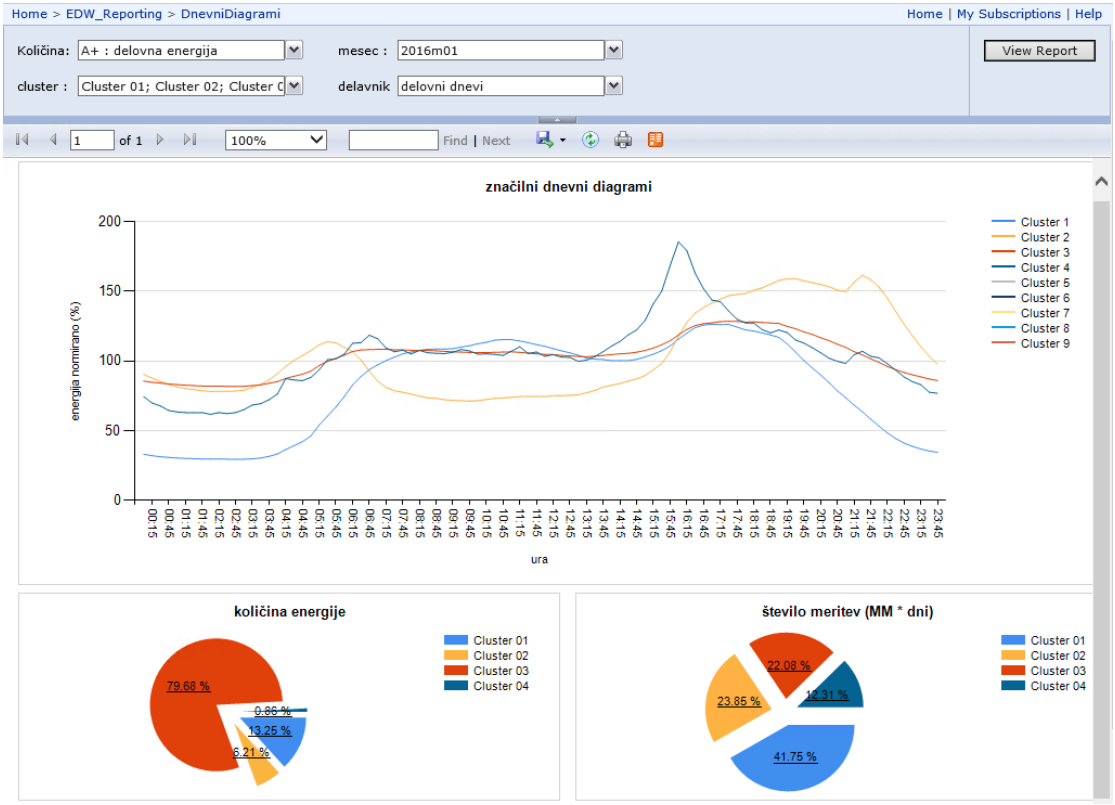
OLAP kocka, katere izdelavo smo opisali v prejšnjih točkah tega poglavja, nam lahko služi kot dobra osnova za analize, povezane z odkrivanjem odtekanja prihodkov.

Kot že rečeno (v točki 4.3), lahko za različne potrebe uporabimo različne odjemalce:

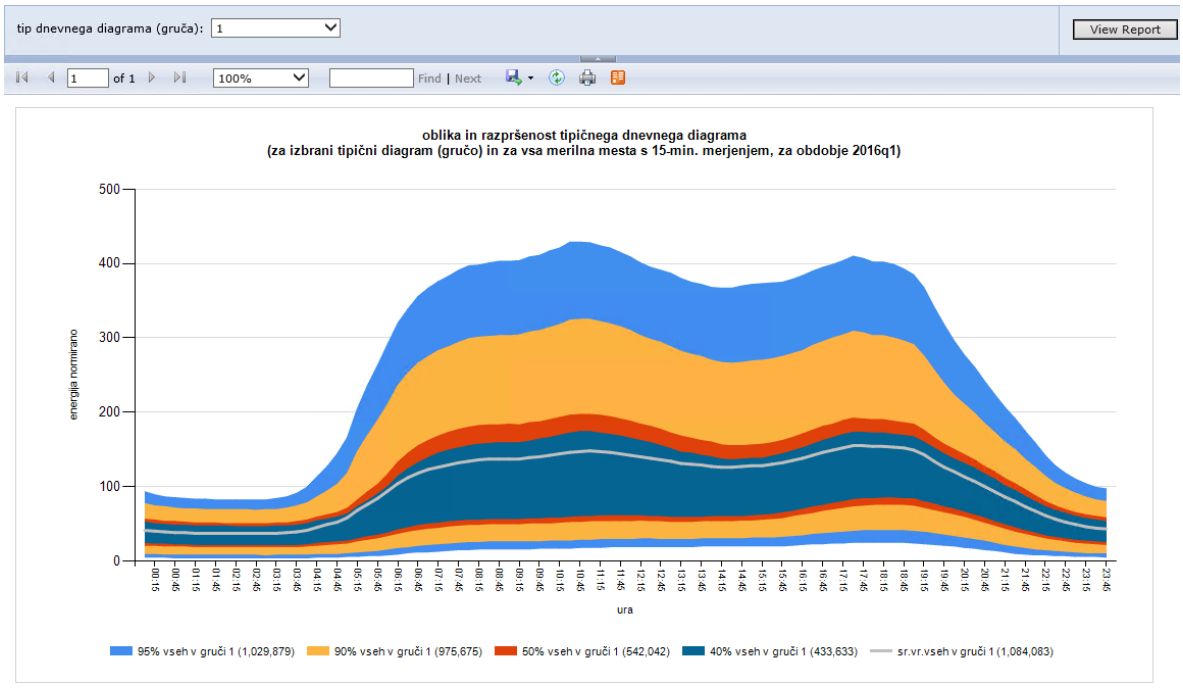
- SSRS (poročevalske storitve strežnika SQL) - za poročila ustaljene oblike. Razvijalci poročil jih po dogovoru pripravijo in objavijo na poročilnem strežniku. Parametri omogočajo uporabniku interaktivnost.
- Excel - To možnost običajno uporabijo naprednejši uporabniki – analitiki za poglobljene in ad hoc analize.
- Spletne storitve Excel (Excel Services) - gre za razširitev portala SharePoint z aplikacijsko storitvijo, ki omogoča nalaganje, preračunavanje in prikaz delovnih zvezkov Excel na spletnem strežniku. Je 'hibridna' možnost, ki združuje prednosti prvih dveh. Napredni uporabniki z Excelom pripravijo analize, ki so zanimive za skupino in jih objavijo na portalu SharePoint. Ostali odprejo poročilo s spletnim brskalnikom, če želijo, pa tudi z Excelom, kjer lahko poročilo spreminjajo.
- Odjemalec analitičnih storitev je lahko tudi poslovna aplikacija, kateri omogočimo poizvedovanje po analitski strukturi. To je lahko tudi periodična obdelava, ki ob izpolnjenem določenem pogoju poizvedbe sproži obveščanje ali kakšno drugo akcijo.

Za demonstracijo smo pripravili nekaj poročil. Poročila, prikazana na slikah 4.33 do 4.36, so izdelana kot spletna poročila, objavljena s poročevalskimi storitvami strežnika SQL (SSRS).

S poročiloma na slikah 4.33 in 4.34 lahko uporabniki hitro in enostavno pregledajo lastnosti značilnih dnevnih diagramov, ki so nastali s podatkovnim rudarjenjem, kot je opisano v točki 4.4. Z deleži so prikazane tudi lastnosti množice posamičnih dnevih diagramov, ki so razvrščeni v določeno skupino.

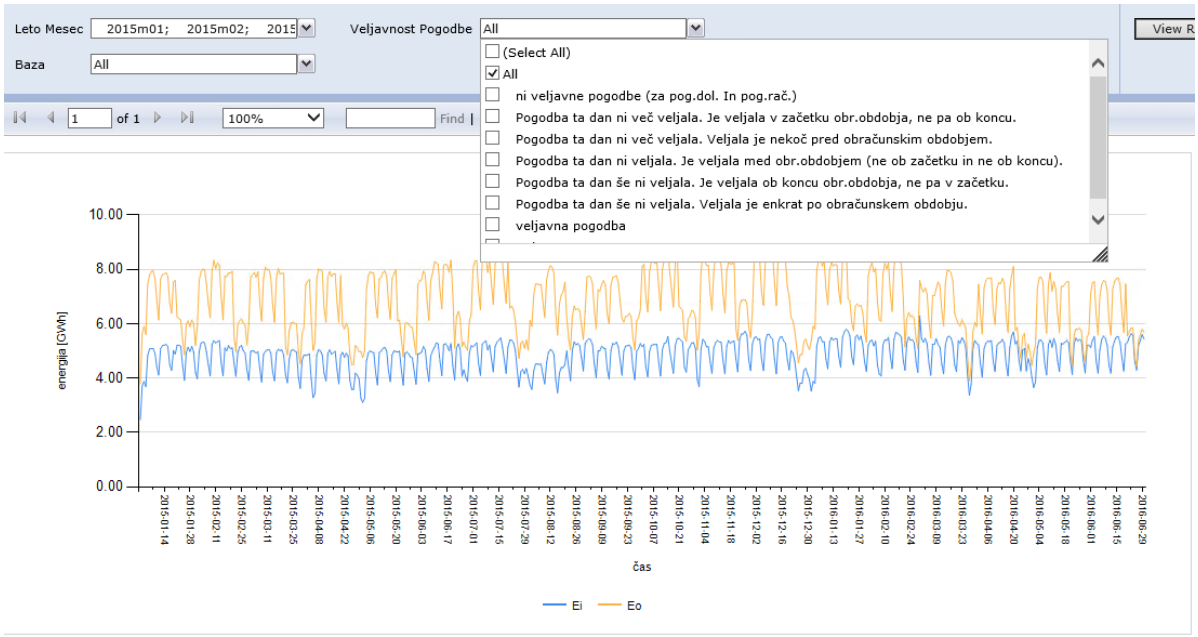


Slika 4.33: Interaktivno spletno poročilo za analizo oblike značilnih dnevnih diagramov

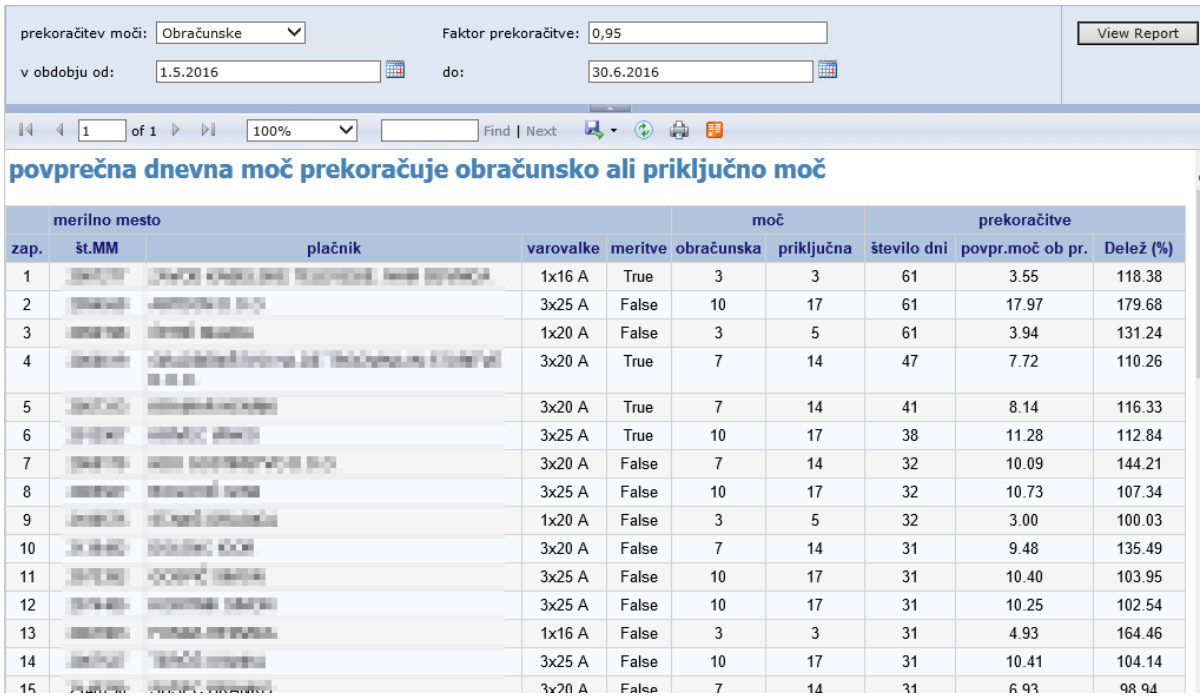


Slika 4.34: Interaktivno spletno poročilo za analizo razpršenosti razvrščenih dnevnih diagramov

S poročilom, ki ga prikazuje slika 4.35, lahko za izbrano obdobje enega ali več mesecev primerjamo gibanje obračunane energije in energije izmerjene v sistemu AMI, razdeljene po dneh.



Slika 4.35: Interaktivno spletno poročilo s prikazom diagrama obračunane energije in energije, izmerjene z naprednim sistemom merjenja (AMI)



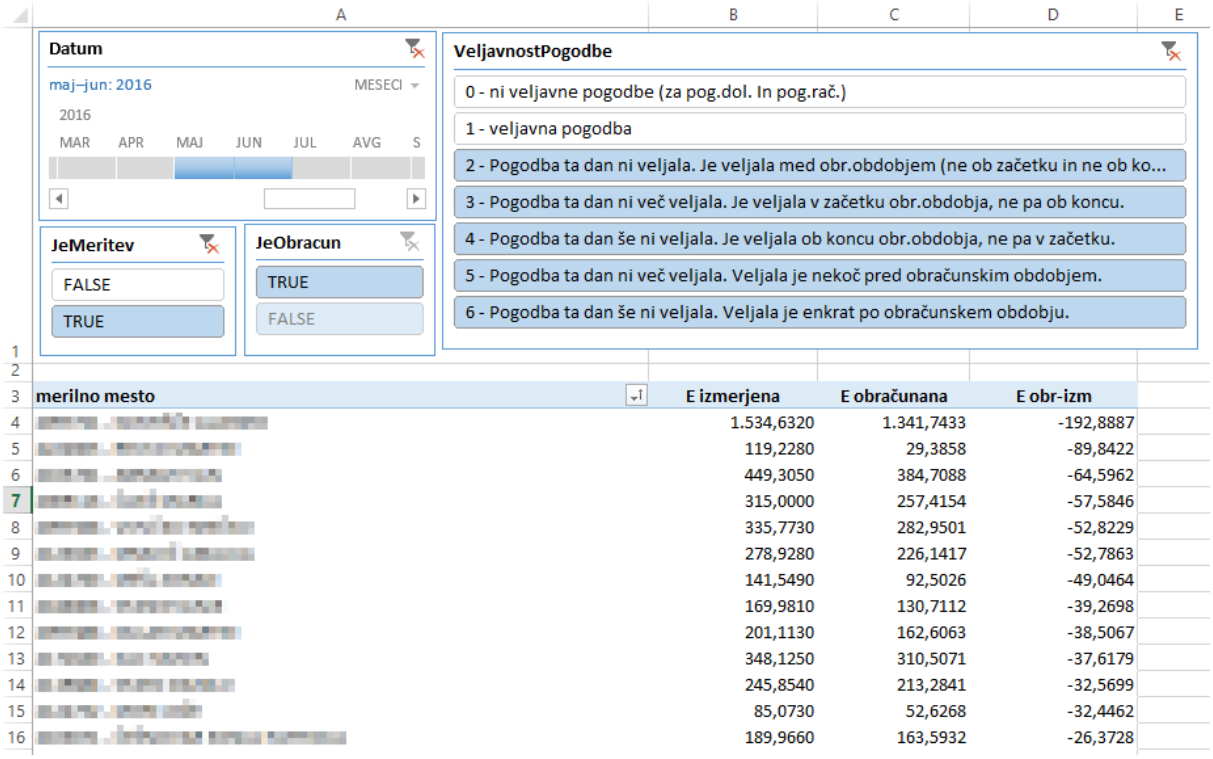
	merilno mesto				moč		prekoračitve		
zap.	št.MM	plačnik	varovalke	meritve	obračunska	priključna	število dni	povpr.moč ob pr.	Delež (%)
1	[redacted]	[redacted]	1x16 A	True	3	3	61	3.55	118.38
2	[redacted]	[redacted]	3x25 A	False	10	17	61	17.97	179.68
3	[redacted]	[redacted]	1x20 A	False	3	5	61	3.94	131.24
4	[redacted]	[redacted]	3x20 A	True	7	14	47	7.72	110.26
5	[redacted]	[redacted]	3x20 A	True	7	14	41	8.14	116.33
6	[redacted]	[redacted]	3x25 A	True	10	17	38	11.28	112.84
7	[redacted]	[redacted]	3x20 A	False	7	14	32	10.09	144.21
8	[redacted]	[redacted]	3x25 A	False	10	17	32	10.73	107.34
9	[redacted]	[redacted]	1x20 A	False	3	5	32	3.00	100.03
10	[redacted]	[redacted]	3x20 A	False	7	14	31	9.48	135.49
11	[redacted]	[redacted]	3x25 A	False	10	17	31	10.40	103.95
12	[redacted]	[redacted]	3x25 A	False	10	17	31	10.25	102.54
13	[redacted]	[redacted]	1x16 A	False	3	3	31	4.93	164.46
14	[redacted]	[redacted]	3x25 A	False	10	17	31	10.41	104.14
15	[redacted]	[redacted]	3x20 A	False	7	14	31	6.93	98.94

Slika 4.36: Interaktivno spletno poročilo z merilnimi mesti, ki prekoračujejo obračunsko ali priključno moč

Na sliki 4.36 je prikazano poročilo v obliki preglednice, s katerim odkrijemo merilna mesta, ki v izbranem obdobju prekoračujejo obračunsko ali priključno moč (odvisno od izbire) za določen odstotek. Zapisi so razvrščeni padajoče po številu dni prekoračevanja. Tisti pri vrhu, ki konstantno prekoračujejo eno od izbranih moči, imajo skoraj zagotovo vgrajene varovalke za večji tok, kot je zanje predpisano. Še posebej verjetno je to pri odjemalcih s pametnim števcem, saj so zanje prikazane dejanske prekoračitve, medtem ko so pri odjemalcih s klasičnimi števci prikazane prekoračitve glede na ocenjene dnevne količine.

To poročilo je zelo koristno za sprotno odkrivanje odtekanja prihodkov. Odjemalci, ki prekoračujejo obračunsko moč in so si skoraj zagotovo nedovoljeno vgradili prevelike varovalke, bi morali povečati svojo obračunsko moč, kar bi zanje pomenilo večji mesečni strošek omrežnine za obračunsko moč, za elektrodistribucijsko podjetje pa večji prihodek. Tisti, ki pa prekoračujejo celo priključno moč, bi morali zaprositi za novo soglasje in po izdaji soglasja ob sklenitvi nove pogodbe o priključitvi tudi plačati enkratno razliko omrežnine za priključno moč. Nato se jim spremeni tudi obračunska moč in s tem mesečni strošek omrežnine za obračunsko moč.

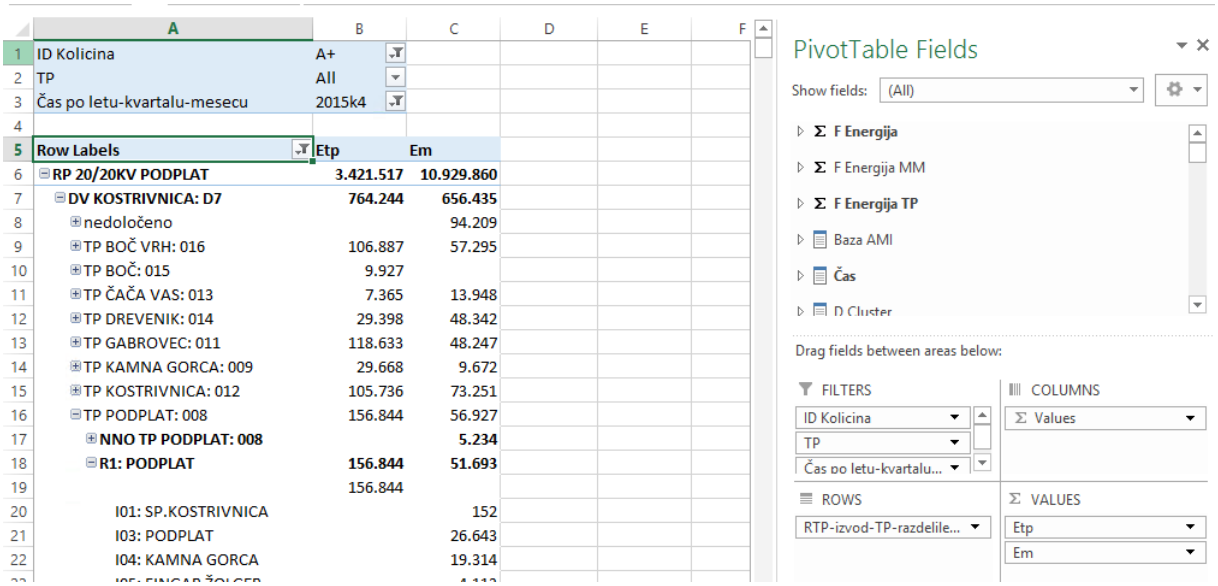
Poročili, prikazani na slikah 4.37 in 4.38 sta izdelani z Excelom. Prvo poročilo (slika 4.37) v obliki preglednice za merilna mesta prikazuje njihovo porabo, izmerjeno s sistemom naprednega merjenja (AMI), v primerjavi s količino energije, ki je bila v tem obdobju obračunana. Atribut 'JeMeritev', po katerem lahko omejimo prikazane zapise, ima vrednost 'resnično' za tista merilna mesta, ki imajo nameščen pametni števec, vključen v sistem AMI. Atribut 'JeObracun' ima vrednost 'neresnično' za tista merilna mesta, ki za izbrano obdobje niso imela obračuna. Posamično merilno mesto lahko ima v izbranem obdobju več obračunov. Vsak obračun je izdelan za obdobje. Tudi pogodba o dostopu na merilnem mestu ima obdobje veljavnosti. Glede na prekrivanje teh dveh obdobij lahko pride do različnih situacij veljavnosti pogodbe, kar podaja parameter 'VeljavnostPogodbe', po katerem tudi lahko zožujemo seznam. Kombinacije s kakorkoli neveljavno pogodbo in hkrati izmerjeno energijo takoj sprožijo sum na nepravilnosti. Prav tako sprožijo sum na nepravilnosti kombinacije z veljavno pogodbo in izmerjeno energijo, a brez obračuna. Vendar so pogoste in čisto regularne zadeve, ko gre za zamenjavo lastništva ali najem prostorov in ob tem za poračun za obdobje ali del obdobja, ko pogodba ni bila več ali še veljavna. Vsekakor pa so zapisi, ki jih na tak način 'ulovimo' kot sumljive, dobri kandidati za preverjanje v transakcijskem sistemu.



merilno mesto	E izmerjena	E obračunana	E obr-izm
[redacted]	1.534,6320	1.341,7433	-192,8887
[redacted]	119,2280	29,3858	-89,8422
[redacted]	449,3050	384,7088	-64,5962
[redacted]	315,0000	257,4154	-57,5846
[redacted]	335,7730	282,9501	-52,8229
[redacted]	278,9280	226,1417	-52,7863
[redacted]	141,5490	92,5026	-49,0464
[redacted]	169,9810	130,7112	-39,2698
[redacted]	201,1130	162,6063	-38,5067
[redacted]	348,1250	310,5071	-37,6179
[redacted]	245,8540	213,2841	-32,5699
[redacted]	85,0730	52,6268	-32,4462
[redacted]	189,9660	163,5932	-26,3728

Slika 4.37: Poročilo za primerjavo izmerjene (AMI) in obračunane energije, izdelano v Excelu

Poročilo, prikazano na sliki 4.38, predstavlja povsem generično vrtilno tabelo, ki za vir podatkov uporabi analitične storitve strežnika SQL (SSAS). V prikazanem primeru sta kot vrednosti, s katerima izvajamo analizo, izbrani količini energije iz dveh različnih tabel dejstev: Em - iz meritev, zbranih s sistemom AMI na merilnih mestih odjemalcev in Etp – iz meritev, zbranih s sistemom AMI, s števcev, vgrajenih v transformatorskih postajah na nizkonapetostnih izvodih. Vrednosti iz teh dveh tabel dejstev povezujemo in primerjamo preko dimenzije 'omrežje', ki jo prikazujemo v vrsticah vrtilne tabele v obliki hierarhije 'RTP-izvod-TP-razdelilec', opisane v točki 4.2.6. S takšnim poročilom zlahka odkrivamo razlike med energijo, ki jo izmerimo na izvodu iz transformatorske postaje in vsoto energije, ki so jo porabili odjemalci, priključeni na to transformatorsko postajo. Kadar je energija na izvodu transformatorske postaje večja, kot vsota količin pri odjemalcih, bi v primeru, ko so v sistem AMI vključeni vsi odjemalci na tem izvodu pomenilo, da je nekaj energije in prihodka odteklo mimo števcev. Tehnične izgube običajno znašajo manj kot 5%.



	A	B	C
1	ID Kolicina	A+	
2	TP	All	
3	Čas po letu-kvartalu-mesecu	2015k4	
4			
5	**Row Labels**	**Etp**	**Em**
6	**RP 20/20KV PODPLAT**	**3.421.517**	**10.929.860**
7	**DV KOSTRIVNICA: D7**	**764.244**	**656.435**
8	nedoločeno		94.209
9	TP BOČ VRH: 016	106.887	57.295
10	TP BOČ: 015	9.927	
11	TP ČAČA VAS: 013	7.365	13.948
12	TP DREVENIK: 014	29.398	48.342
13	TP GABROVEC: 011	118.633	48.247
14	TP KAMNA GORCA: 009	29.668	9.672
15	TP KOSTRIVNICA: 012	105.736	73.251
16	TP PODPLAT: 008	156.844	56.927
17	**NNO TP PODPLAT: 008**		**5.234**
18	**R1: PODPLAT**	**156.844**	**51.693**
19		156.844	
20	I01: SP.KOSTRIVNICA		152
21	I03: PODPLAT		26.643
22	I04: KAMNA GORCA		19.314

Slika 4.38: Poročilo z vrtilno tabelo, izdelano z Excelom, s podatkovnim virom SSAS (OLAP)

Tovrstno, seveda še precej dodelano poročilo, je lahko zelo učinkovito orodje pri preprečevanju odtekanja prihodkov. Predpogoj, da bo lahko učinkovito je, da predhodno odpravimo težave s kakovostjo podatkov, opisane v točkah 4.2.3, 4.2.4, 4.2.6 in 4.2.7.

5 Sklepne ugotovitve

Preprečevanje odtekanja prihodkov zagotovo ni najpomembnejša poslovna strategija. Pomembno pa je, da sadove opravljenega dela žanjemo skozi prihodek in jih ne pustimo odtekati, saj nam bo sicer zmanjkalo sredstev za uresničevanje tistih bolj vizionarskih strategij. Pri tem smo lahko veliko bolj uspešni, če aktivnosti zastavimo sistematično. Zelo nam lahko pomagajo analitična informacijska orodja. V dobi velikih količin podatkov tudi poslovno področje distribucije električne energije ni izjema. Z velikimi količinami podatkov se v tem okolju srečujemo predvsem zaradi uvedbe naprednih sistemov merjenja s pametnimi števci. Izziv pa ni le količina podatkov, ampak tudi heterogenost sistemov, zaradi česar je za povezovanje in navzkrižno analizo potrebno vložiti nekaj več napora.

S praktično implementacijo smo dokazali, da je s klasičnimi pristopi možno obvladovati množične podatke, zbrane v naprednih sistemih merjenja porabe električne energije. Iz množice podatkov smo izluščili informacije tako, da smo podatke prečistili, preoblikovali in povezali v podatkovno skladišče ter iz njega sestavili analitično strukturo. Z raziskovanjem vzorcev v izmerjenih količinah električne energije smo z uveljavljenimi metodami strojnega učenja uspešno odkrivali znanje, skrito v teh podatkih in ugotavljali, kakšne so značilne oblike dnevne porabe in kakšno je najprimernejše število različnih tipov oblik, ki jih lahko identificiramo.

Preizkusili smo različne algoritme za razvrščanje v skupine. Njihovo uspešnost smo medsebojno primerjali s preverjanjem razpršenosti in ocenjevanjem smiselnosti dobljenih razvrstitev. Učinkovitost algoritmov smo primerjali z merjenjem časa procesiranja. Kot najboljšo smo izbrali neskalabilno različico metode maksimiranja pričakovanj (EM). S ponavljanjem indukcije modelov za različno število skupin in razvrščanja vzorcev vanje smo prišli do zaključka, da je najprimerneje vzorce razvrstiti v štiri skupine za vsak letni čas oz. kvartal in ločeno za delovnike in dela proste dni. S po dvema modeloma za vsak kvartal smo vse prečiščene in normirane vzorce razvrstili v skupine podobnih dnevnih diagramov, katerih srednje vrednosti predstavljajo značilne dnevne diagrame.

Z iskanjem korelacij med lastnostmi merilnih mest oz. njihovih pogodb o dostopu, časom vzorca ter pripadnosti značilni skupini, smo nadaljevali odkrivanje znanja. Tokrat smo uporabili več metod klasifikacije in predikcije. Med zelo podobno uspešnimi smo kot najprimernejšo izbrali metodo z odločitvenimi drevesi. S po dvema modeloma za vsak kvartal smo v znane ciljne skupine na osnovi ocenjene dnevne porabe uvrstili vsa tista merilna mesta, za katera ni na voljo dnevnih diagramov s 15-minutnimi porabami. Uspešnost tega uvrščanja ocenjujemo na dokaj nizko, kar je tudi pokazala faza validacije s testnimi množicami in z navzkrižnim preverjanjem. To pripisujemo dejstvu, da lastnosti odjemalčeve pogodbe in količina porabljene energije še zdaleč niso edine spremenljivke, ki določajo dinamiko porabe. Domnevamo, da so pomembni dejavniki tudi temperatura, osvetljenost, velikost in energetska učinkovitost stavb, število prebivalcev v stavbi in podobno. Žal pa teh podatkov nismo mogli pridobiti, da bi lahko preverili njihov vpliv na uspešnost in izboljšali rezultate. Kljub temu smo, ponovno z metodami

strojnega učenja prišli do pravil, katere lastnosti odjemalca vplivajo na obliko porabe. Ta pravila so zaradi navedenih omejitev bolj šibka in pomanjkljiva.

Tako smo za vsa merilna mesta v omrežju pridobili dnevne količine in dinamike porabe električne energije, ki so za nekatera natančno izmerjene, za druga ocenjene in predvidene. Skupaj z ostalimi podatki, urejeno zbranimi v podatkovnem skladišču ter predpripravljenimi za analizo v OLAP kocki, ta celota predstavlja zelo dobro osnovo za odkrivanje odtekanja prihodkov v elektrodistribuciji. To potrjujejo vzorčna poročila, ki smo jih izdelali na osnovi te podatkovne strukture in jih predstavili v zadnjem delu naloge.

Celoten proces priprave analitične strukture za podporo preprečevanja odtekanja prihodkov, ugotovitve, ki izvirajo iz tega procesa in izdelana rešitev sama, prinašajo še stranske rezultate, ki nam lahko zelo koristijo tudi na drugih področjih. Na ta način bolje spoznamo obnašanje svojih odjemalcev v odvisnosti od zunanjih dejavnikov, hitreje zaznamo trende in spremembe navad in posledično lahko bolje in pravočasno odreagiramo s prilagoditvami omrežja. Tako lahko tudi za nekoliko dolgoročnejše napovedi porabe energije izdelamo veliko boljše modele, za katere verjamemo, da bodo dali tudi boljše rezultate.

Uspešneje lahko izvajamo tudi analize pretokov moči in padcev napetosti ter kratkostičnih razmer v omrežju, saj lahko iz dobro organiziranega skladišča hitro in enostavno ponudimo podrobne in natančne vhodne podatke. S tem prispevamo k boljši optimizaciji sistema.

Glavne omejitve, s katerimi smo se srečali pri izvedbi naloge, so vezane na kakovost podatkov in na nerazpoložljivost nekaterih podatkov, npr. o temperaturah. Ravno s kakovostjo podatkov je zelo povezan dobršen del pristopa preprečevanja odtekanja prihodkov. Kot ena od temeljnih usmeritev mora biti skrb za odkrivanje in sprotno odpravljanje vsakršnih anomalij v podatkih. Ključno je, da so različni sistemi, ki sodelujejo v celotnem poslovnem procesu, že v osnovi povezani tako, da zagotavljajo konsistentnost.

Naloga pušča še veliko priložnosti za nadaljnje delo.

Pri izvedbi praktične rešitve za podporo preprečevanju odtekanja prihodkov bi veljalo razmisliti oz. nadaljevati z:

- razširitvami podatkovnega skladišča s podatki o:
  - financah (zneskih),
  - podatkih iz faze finančne realizacije in izterjave,
  - ostalih dogodkih, ki jih zajemajo pametni števci,
  - opravljenih kontrolah merilnih mest,
  - opravljenih servisnih posegih (na zahtevo stranke),
  - evidentiranih odkritih krajah električne energije in s tem povezanih goljufijah,
  - ostalih storitvah, ki jih opravlja elektrodistribucijsko podjetje strankam,
  - vremenu, temperaturi, osvetljenosti,
- izdelavo dodatnih poročil za iskanje nepravilnosti v podatkih,
- izdelavo osrednje 'kontrolne plošče' z zbirnimi poročili in ključnimi kazalniki,
- posebno pozornost v analizi posvetiti proizvodnji električne energije,

- pri odkritih krajah raziskati značilne vzorce obnašanja in to znanje vgraditi v postopke odkrivanja,
- itd.

Kot smo napovedali uvodoma, v tej nalogi nismo veliko pozornosti posvečali poslovnim procesom za preprečevanje odtekanja prihodkov, kar pa ne pomeni, da ti niso pomembni. Za urejeno in zrelo podjetje je ključno, da vzpostavi dobro definirane postopke ter metriko s kazalniki za spremljanje in izboljševanje njihove uspešnosti. Dobra orodja, ki temeljijo na napredni analitiki in odkrivanju znanja v velikih množicah podatkov, so nujen predpogoj za uspešnost, a sama po sebi še ne zadoščajo.

Kazalo slik

Slika 2.1: Shematski prikaz udeležencev na trgu z električno energijo (vir: http://www.agen-rs.si/udelezenci-na-trgu-z-elektricno-energijo) 7

Slika 2.2: Opredelitev konceptov podatek, informacija, znanje, modrost in razmerja med njimi (povzeto po: https://en.wikipedia.org/wiki/DIKW_Pyramid) 14

Slika 3.1: Primer množice s 5 2-razsežnimi elementi in dendrogram njihovega postopnega združevanja 21

Slika 3.2: Prikaz razvrščanja elementov po metodi K-voditeljev (za K=3) 22

Slika 4.1: Gradniki poskusnega sistema 29

Slika 4.2: Model podatkovnega skladišča EDW 31

Slika 4.3: Pomen posameznih bitov v atributu 'status' pri meritvah pametnega števca 33

Slika 4.4: Poenostavljen entitetni diagram konceptov, ki jih združene (denormalizirano) zapišemo v dimenzijo 'pogodba o dostopu' 38

Slika 4.5: Obdobje veljavnosti zapisa o pogodbi o dostopu 39

Slika 4.6: Preslikava omrežja iz BTP v dimenzijo 'omrežje' (sredstva različnih vrst, upoštevaje navedene vrste medsebojnih povezav, zapišemo v dimenzijsko tabelo po nivojih) 41

Slika 4.7: ETL postopek za inicialno polnjenje podatkovnega skladišča 44

Slika 4.8: ETL postopek za redno (dnevno) dodajanje in posodabljanje podatkovnega skladišča 45

Slika 4.9: Vmesnik SSDT za določanje strukture OLAP kocke 50

Slika 4.10: Vmesnik SSDT za urejanje dimenzije 50

Slika 4.11: Vmesnik SSDT za določanje uporabe dimenzij v povezavi s skupinami mer 51

Slika 4.12: Primer analize OLAP podatkov z odjemalcem Excel 52

Slika 4.13: Pregled lastnosti modela rudarjenja v diagramu skupin 55

Slika 4.14: Pregled lastnosti modela rudarjenja v obliki profilov skupin 56

Slika 4.15: Del generičnega pregledovalnika lastnosti modela podatkovnega rudarjenja 56

Slika 4.16: Urejanje in izvajanje prediktivne poizvedbe z orodjem SSDT 57

Slika 4.17: Vsote evklidskih razdalj testne množice v odvisnosti od števila skupin, za različne metode razvrščanja 58

Slika 4.18: Čas procesiranja modela in čas razvrščanja za različne metode združevanja 58

Slika 4.19: Populacija skupin pri razvrščanju podatkov za delovnike prvega kvartala 2016 z različnimi modeli in za različno število skupin 59

Slika 4.20: SSIS paket za postopek razvrščanja 60

Slika 4.21: Tipični dnevni diagrami posamične skupine ob delovnikih, za kvartale..... 61

Slika 4.22: Skupne količine energije (zgoraj) in število zapisov (spodaj) v skupinah, ločeno za delovne (levo) in dela proste dni (desno), po kvartalih 62

Slika 4.23: Deleži populacij skupin po kvartalih, ločeno za delovnike in dela proste dni 62

Slika 4.24: Deleži količin porabe (levo) in števila elementov (desno) v skupinah, v povprečjih za vse kvartale in združeno za delovne in dela proste dni 62

Slika 4.25: Število merilnih mest v odvisnosti od deležev pogostosti pojavljanja v svoji prvi, drugi in tretji najpogostejši skupini 63

Slika 4.26: Število merilnih mest, opredeljenih s posamezno skupino, s kombinacijo dveh ali s kombinacijo treh skupin tipičnih dnevnih porab 64

Slika 4.27: Struktura in modeli za podatkovno rudarjenje, izdelani v razvojnem okolju SSDT 65

Slika 4.28: Vmesnik SSDT za nastavitev parametrov za izdelavo modela odločitvenega drevesa 66

Slika 4.29: Pregled vpliva posameznih atributov na določitev ciljne spremenljivke 66

Slika 4.30: Pregledovalnik odločitvenega drevesa v okolju SSDT 67

Slika 4.31: Odzivni diagram za izdelane modele..... 68

Slika 4.32: Grafična ponazoritev celotnega postopka razvrščanja merilnih mest in njihove dnevne porabe..... 70

Slika 4.33: Interaktivno spletno poročilo za analizo oblike značilnih dnevnih diagramov 72

Slika 4.34: Interaktivno spletno poročilo za analizo razpršenosti razvrščenih dnevnih diagramov 72

Slika 4.35: Interaktivno spletno poročilo s prikazom diagrama obračunane energije in energije, izmerjene z naprednim sistemom merjenja (AMI)..... 73

Slika 4.36: Interaktivno spletno poročilo z merilnimi mesti, ki prekoračujejo obračunsko ali priključno moč..... 73

Slika 4.37: Poročilo za primerjavo izmerjene (AMI) in obračunane energije, izdelano v Excelu 75

Slika 4.38: Poročilo z vrtilno tabelo, izdelano z Excelom, s podatkovnim virom SSAS (OLAP) 76

Literatura

1 K. Baamann, »Data Quality Aspects Of Revenue Assurance«, 2007 International Conference on Information Quality, MIT ICIQ, Cambridge USA, November 2007

2 A.S. Bassan, D.Sarkar, Mastering SQL Server 2014 Data Mining, Packt publishing, 12-2014

3 G. Berginc, S. Gašperič, Razvrščanje obremenitvenih diagramov z verjetnostnim nevronskim omrežjem, 6. konferenca slovenskih elektroenergetikov CIRED, Portorož, 2003

4 C.M.Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, 2006

5 P. S. Bradley, U. Fayyad, C. Reina, Scaling EM (Expectation-Maximization) Clustering to Large Databases, Microsoft Research, 1998

6 S. Gašperič, D. Gerbec: Izdelava nadomestnih obremenitvenih diagramov za slovensko distribucijsko omrežje, Univerza v Ljubljani, Fakulteta za elektrotehniko, maj 2004

7 J.P. Gouveia ∗ , J. Seixas, »Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys«, Energy and Buildings, Volume 116, March 2016 (666-676)

8 J. I. Guerrero, C. León, I. Monedero, F. Biscarri, J. Biscarri, »Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection«, Knowledge-Based Systems, Volume 71, November 2014, (376 – 388)

9 J.Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufman publications, 2006

10 D. Hand, H. Mannila, P. Smyth, Principles of data mining, A Bradford Book The MIT Press, 2001

11 M. Hayn , V. Bertsch, W. FichtnerElectricity, »Load profiles in Europe: The importance of household segmentation«, Energy Research & Social Science Volume 3, September 2014, (30 – 45)

12 N. Ilc, "Primerjava metod za razvrščanje vzorcev v gruče", Diplomska naloga na univerzitetnem študiju, UL, FRI, 2009

13 F. McLoughlin, A. Duffy, M. Conlon, »A clustering approach to domestic electricity load profile characterisation using smart metering data«, Applied Energy, Volume 141, March 2015, (190 – 199)

14 J. Nagi et al., (2010) “Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines”, IEEE Transactions on Power Delivery, Vol. 25, N. 2, Apr 2010.

15 C. C.O. Ramos, A.N. Souza, G. Chiachia, A.X. Falcão, J. P. Papa, »A novel algorithm for feature selection using Harmony Search and its application for non-technical losses detection«, Computers and Electrical Engineering, Volume 37, Issue 6, November 2011 (886–894)

16 M. Riveira, R. Johansson, A. Karlsson, »Modeling and analysis of energy data: state-of-the-art and practical results from an application scenario«, article, University of Skövde , Skövde, 2011

17 Program razvoja pametnih omrežij v Sloveniji, del I: Distribucijsko omrežje, SODO d. o. o., Fakulteta za elektrotehniko Univerze v Ljubljani in Elektroinštitut Milan Vidmar, Ljubljana, 2012

18 P. Šaponja, "Odkrivanje skupin s pomočjo argumentiranega strojnega učenja", Magistrsko delo, UL, FRI, 2015

19 Revenue Assurance Overview – Technical Report TR131
dostopno na: https://www.tmforum.org/resources/standard/tr131-revenue-assurance-overview-v2-4-1/

20 Revenue Assurance Guidebook, TeleManagement Forum, 2014
dostopno na: https://www.tmforum.org/resources/standard/gb941-main-revenue-assurance-guidebook-v3-5-1/

21 Akt o metodologiji za določitev regulativnega okvira in metodologiji za obračunavanje omrežnine za elektrooperaterje, Uradni list RS, št. 66/2015

22 Sistemska obratovalna navodila za distribucijsko omrežje električne energije (SONDO), Uradni list RS, št. 41/2011

23 J. Verbeek, N. Vlassis, J. Nunnink, A variational EM algorithm for large-scale mixture modeling, HAL, 2011, Dostopno na: https://hal.inria.fr/inria-00321486v1

24 Wand,Y and Wang, R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations, Communications of the ACM, 39(11), 1996, pp. 86-95

25 I.H.Witten, F.Eibe, M.A.Hall, Data mining: practical machine learning tools and techniques—3rd Edition, Elsevier, Morgan Kaufmann publications, 2011